THE RELATIONSHIP OF WORD POWER AND NOISE-MASKED GAIN FUNCTION PARAMETERS TO THE INTELLIGIBILITY OF WORD SETS

A THESIS

Presented to the

Faculty of the Graduate Division

by

George Boltz Hawthorne, Jr.

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in the School of Electrical Engineering

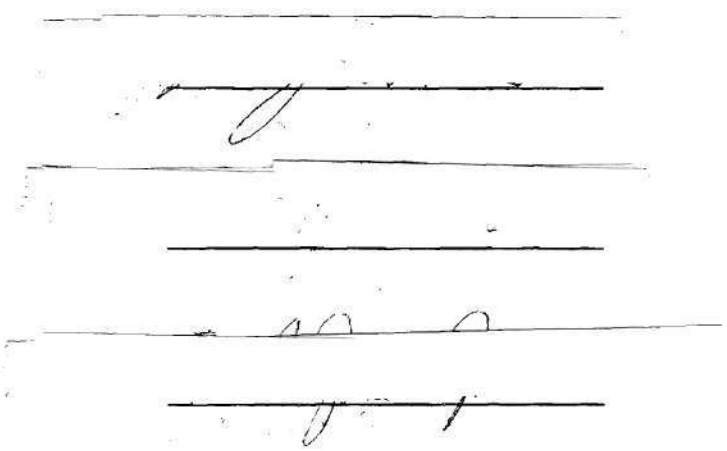
Georgia Institute of Technology

December, 1962

55
12R

THE RELATIONSHIP OF WORD POWER AND NOISE-MASKED GAIN FUNCTION PARAMETERS TO THE INTELLIGIBILITY OF WORD SETS

Approved:



Date Approved by Chairman: Jun 21, 1963

DEDICATION

This thesis is dedicated to my wife Ethel, whose patience and encouragement made its completion possible.

ACKNOWLEDGMENTS

. The author is indebted to his thesis advisor, Dr. B. J. Dasher, for his generous assistance and encouragement throughout the course of this investigation. The author also gratefully acknowledges the equipment and facilities made available to him by Georgia Tech's School of Electrical Engineering and Engineering Experiment Station. Staff members of the latter, particularly in the Rich Electronic Computer Center, the Communications Branch, and the Special Problems Branch, are due thanks for their support during the experimental phase of the study.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Continued)

LIST OF ILLUSTRATIONS (Continued)

SUMMARY

Conventional articulation test procedures as described, for example, by Egan have been widely used to measure the intelligibility of speech and the performance of speech communication systems. Such tests consist basically of passing a set of speech items, such as syllables, words, or sentences, through a transmission channel and presenting the channel output to a group of human subjects (listeners). The percentage or fraction of speech items correctly identified by the listeners is known as the articulation or articulation score. In the case where words are used as speech items, this quantity is referred to as word articulation or word score. The data from such tests are conventionally presented in the form of an articulation curve, this being a plot of word score versus some test variable. The test variable is usually some transmission property of the channel, such as bandwidth or amount of noise added to the speech. In this study, the intelligibility of the speech was degraded by the addition of masking noise, and the test variable was signal/noise ratio.

The articulation curve may be viewed as representing the intelligibility characteristics possessed by the word set under the conditions of the test. While such a curve is useful for many purposes, the information it contains concerns only the behavior of the word set as a whole. Such a curve contains no information concerning the individual words, and cannot be used to predict the intelligibility of either a given word or of a subset of the original word set.

The main object of this study is to relate noise-masked articulation score, as conventionally defined, to a set of basic speech and noise parameters defined for individual words. Suitable parameters are shown to be, for a given noise, the power, the noise-masked threshold of intelligibility, and the width of the threshold region. Experimentally-determined values of these parameters are presented, along with their distributions, for a limited number of monosyllabic words imbedded in random white noise and for a particular combination of speaker and listeners. It is shown that the threshold parameter can be obtained from the noise-masked gain function, which in turn can be obtained from conventional articulation test data. It is shown that word power and noise power can be obtained from measurements of energy and time duration for individual words and for word-length bursts of noise. The equipment and techniques for measuring energy and time duration are described.

By presenting conventional test data in a novel form, attention is directed to the threshold action occurring as a word "emerges," intelligibility-wise, from masking noise. The existence and sharpness of these thresholds is demonstrated for the set of forty test words and for each of the three listeners, and the nature of the threshold region is illustrated by means of noise-masked gain functions. The general nature of these functions, which are essentially articulation curves for individual words, is examined, and two describing parameters of such functions, namely, threshold and spread, are defined and tabulated. The set of tabulated values is used to calculate and plot articulation curves, which are then compared to experimental curves. From the close agreement of calculated and experimental curves, it is concluded that the set of

parameter values contains essentially the same information contained in the conventional articulation curve, thus exhibiting the basic role of such parameters in determining the intelligibility characteristics of word sets.

In addition to describing the intelligibility characteristics of sets of words, these parameters exhibit the characteristics of individual words, showing the variations from word to word and from listener to listener. It is shown that these properties make the word parameter approach useful in two ways:

(1) It facilitates the analysis of articulation test results, yielding more information than the conventional type of data and exhibiting the effects of individual words on the articulation curve as well as permitting a word-by-word comparison of results from different listeners.

(2) It makes possible the synthesis of word-sets to obtain a specified articulation curve.

The usefulness of the word parameter approach in synthesizing a desired articulation curve is illustrated for four distinct types of curve, two of which have the same shape but different locations and two of which have the same location but different shapes. In this connection, it is demonstrated that, for a given masking noise, the shape of the articulation curve is determined primarily by the threshold-to-power ratios of the individual words. The effect of the shaping technique on the histograms for threshold and for spread is illustrated for four subsets.

A procedure, with variations, is developed for predicting articulation curves in certain cases. This procedure is based on a mathematical

formulation of the process by which individual words contribute to the articulation curve, and makes use of the word parameters referred to above. Essentially two prediction schemes are outlined, based on different degrees of approximation to the noise-masked gain functions. Curves obtained by the two schemes are compared with each other and to experimentally-obtained curves, for the case where listeners are trained and gain function data is obtained on a master word set and where curves are desired for various subsets of the master set. Under such conditions, the nature of the subset curves depends strongly upon whether or not the listeners are retrained on each subset before making tests. The subset tests in this study were made with no retraining on subset words. That such subset curves can be predicted with good accuracy without making actual tests is demonstrated by comparing predicted curves to experimental curves for four twenty-word subsets.

The discrepancies between predicted and experimental curves are discussed and explained on the basis of partial memorization (inadvertent training) during the tests. A simplified mathematical model of the noise-masked listening process is used to predict the nature of discrepancies arising from this cause, and it is shown that the word sets and/or listeners most closely satisfying the assumptions underlying the model exhibit the predicted discrepancies most markedly.

A number of subsidiary results are given, such as the distribution of noise-masked gain function parameters for various listeners, the use and validity of least squares for fitting straight lines to experimentally determined gain function points, and the distributions of energy, power, and duration for the test words. The listeners are compared on

several bases, and it is shown that listeners may be similar in their responses to a set of words as a whole, while differing significantly in their responses to individual words. It is also shown that the result of averaging predicted curves for several listeners is not significantly different from the result obtained by predicting a single curve from averaged thresholds.

It is concluded that, in the noise masked case, the word parameters used in this investigation play a basic role in determining the articulation curve, and that prediction and curve-shaping schemes based on these parameters are practical and useful. In particular, subset curves are obtainable which are directly comparable one to another, and, when large numbers of such curves are involved, they are obtainable more easily, from the standpoint of maintaining stability of test conditions, than by use of conventional testing techniques. It is further concluded that the word parameter approach provides a consistent and logical basis for examining the variation of intelligibility among listeners and/or words. Finally, two aspects of the word parameter approach which merit further investigation are given.

In an appendix, the detailed preparation of magnetic tapes for use in the tests is described, and, in another appendix, the details of listener training and articulation test procedures are given.

CHAPTER I

INTRODUCTION

Statement of Objectives

The primary objective of this study is to relate the intelligibility characteristics of a set of noise-masked test words to the characteristics of the individual words in the set.

More specifically, the aim is to show that articulation score, as conventionally defined for such words, can be calculated from values of suitably-defined word parameters, and, in certain cases, that the articulation curve is determined by the distribution of one or more of these parameters.

Certain subsidiary objectives, predicated upon the attainment of the primary objective, include the following:

1. Development of a method for predicting articulation curves for subsets of words and/or listeners, using word parameters determined by tests on the master word-set and a team of listeners. This method is to be based on a mathematical formulation of the process by which individual words contribute to the articulation curve, and involves the assumption of no listener training on the subset.

2. Investigation of the nature and quantitative description of the threshold region for individual noise-masked words, and of the validity of using threshold quantities in the prediction scheme of subsidiary objective 1.

3. A study of the way in which word parameters influence the shape

and location of the articulation curve for noise-masked speech, and an experimental measurement of such curves for sets of words chosen on the basis of certain of these parameters.

4. An examination of the word parameters themselves and of their distributions and variations from word to word and from listener to listener.

Summary of Results

A major result of this study is the establishment of certain word parameters as a valid replacement of conventional articulation scores in characterizing the intelligibility of a set of noise-masked words. Examination of the threshold action occurring as a word "emerges" from noise led to the characterization of a word by its "noise-masked gain function," which, together with linear and step-like approximations, was obtained for forty test words and for each of three members of a listening team. Noise-masked threshold, defined as the value of signal/noise ratio at the 50 per cent level of intelligibility, was obtained for each word from its gain function.

Under the conditions for which tests were made, it is shown that the parameters of word power, threshold, and noise-masked gain function spread are sufficient to define the articulation curve, and that these parameters thus contain essentially all the information embodied in the set of articulation scores obtained for a set of words. A comparison of experimental and calculated curves for the three listeners shows a maximum discrepancy, in SN ratios at a given word score, of 1 db, with much smaller discrepancies in the 20-to-80 per cent region of the curves. The maximum discrepancy at the 50 per cent level is 0.25 db. Comparison of

team curves reveals even closer agreement, with no discernible discrepancy at the 50 per cent level.

When the curves are calculated by using only word power and threshold, they are still in good agreement with the experimental curves, particularly at the 50 per cent level where the maximum discrepancy is 0.3 db for individual listeners and 0.0 db for the team.

For a given masking noise, the shape of the curve was found to depend on, and to be closely predictable from, the ratio of threshold to normalized word power. When this fact is used to predict the articulation curves for subsets of words, using data taken for the master set and using subjects trained only on this set, the agreement with experimental curves is good. Considering all three listeners and all four subsets, the discrepancy has a maximum value of 2.7 db, and an average value of 1.3 db, at the 50 per cent level. Somewhat better accuracy is obtained by using noise-masked gain function (NMGF) spreads in the prediction scheme.

The prediction scheme is mathematically formulated on a probabalistic basis, and two modifications of the scheme are presented for the cases of constant noise power and negative-slope NMGF's.

Application of the NMGF parameters of threshold and spread to the shaping of articulation curves is illustrated. Four sets of words were chosen on a basis of these parameters with the following results:

1. Two curves were produced which had the same spread between 20 and 80 per cent points, but which were displaced by approximately four db of signal/noise ratio.

2. Two curves were produced which had no displacement (at the

50 per cent level), but whose 20-to-80 per cent spreads differed by 5.5 db.

Experimentally determined values of energy, time duration, and power for individual words are presented, along with histograms and distribution functions. These indicate, for the monosyllabic test words used, that low-energy words predominate, that time durations are uniformly distributed, and hence that low-power words are predominant. Examination of various word parameters reveals a fairly strong central tendency in the histograms for NMGF threshold and spread, with clear separation of the central peaks of threshold histograms in the case of subsets chosen to give displaced articulation curves. The threshold parameter is shown to have only a weak dependence on NMGF spread. Thresholds are shown to have a definite correlation from listener to listener for two of the three possible listener pairs, while NMGF spreads are shown to have less correlation. Finally, the multiple choices available to a listener, among words having a recognized vowel sound, are shown to be related to the threshold values.

The fact that subset curves can be predicted with good accuracy under certain conditions has two implications of importance in the field of noise-masked articulation testing:

1. A significant reduction in testing time appears attainable, in the case where articulation curves are desired for a large number of subsets having many common words.

2. Subset curves obtained by the prediction scheme are directly comparable, since they are obtained from the same master set of word parameters. Obtaining such comparable curves for large numbers of

subsets is much easier by the methods described than with conventional test methods, mainly because of the difficulty, with conventional methods, of maintaining stable test conditions over extended periods of time.

A final implication of interest is that word sets may be chosen for equal difficulty, i.e., homogeneity of basic audibility or identity of articulation curves, on a basis other than phonetic balance, and that word sets for special applications (such as speech audiometry) may be chosen to yield the desired shape of curve on a basis of the word parameters described.

Background

Articulation testing, as a systematic means of assessing the intelligibility properties of a communication system, had its beginnings in the early years of the twentieth century (1), with much of the earlier work being done at the Bell Telephone Laboratories, beginning about 1919 (2). Over the years, this work has continued, with contributions from such apparently diverse fields as communications engineering, psychology, linguistics, phonetics, mathematics, physics, and certain areas of medicine such as otolaryngology and speech audiometry. Much was done during World War II at Harvard's Psycho-Acoustic Laboratory to standardize articulation testing methods, as exemplified by the work of Egan (3). Egan described the development of a set of phonetically-balanced (PB) word lists containing monosyllabic words whose phoneme content, over each list, closely approximates the distribution of phonemes in the English language. Adding to earlier work at Bell Labs, Egan, and later Hawley (4), described standardized testing methods, while at the Central Institute for the Deaf, various word lists were developed (5) for use in speech audiometry. The

earlier PB words were monosyllables, such as "bar," "jam," and "dill"; lists of spondaic words ("baseball," "cupcake") were also developed at Harvard (PAL Tests number 9 and 14), such "spondees" being dissylables, with equal stress on both syllables. The CID lists are essentially improved versions of the Harvard lists, recorded on disks after correction for varying intensity among words in some cases, and include CID Tests W-1 and W-2 (spondees) and CID Test W-22 (phonetically balanced monosyllables). Various other word, syllable, and sentence lists are used, not only for the study of hearing and the determination of hearing loss for speech, but in the evaluation of speech processing devices and military communications systems (6,7,8,9).

An articulation test, shown in block diagram form in Figure 1, generally consists of transmitting a set of speech items through a communication channel to a team of human subjects (listeners) who attempt to identify the items. The fraction or percentage of items correctly identified is called the articulation score or, simply, the articulation. Various speech units, such as phonemes, syllables, vowels, words, and sentences, are used as test items, and these may be spoken aloud by a

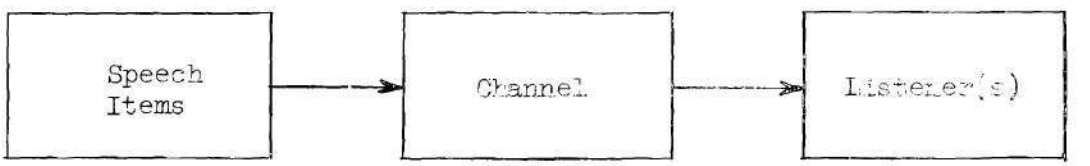


Figure 1. Block Diagram of Articulation Test.

speaker or recorded in some way (as on magnetic tape). The "channel" may be simply a free-field path through air from speaker to listeners, or it may be a complicated electronic communication system which includes transducers to convert the sound-pressure waves of the speaker's voice to electrical form and back again. In the former case, the speech level at any point in the system may be measured in terms of the intensity (rate of sound energy transmission per unit area) or the effective sound pressure (rms value of instantaneous pressure waveform). In the latter case, the analogous quantities of power and rms speech voltage may be measured at any point in the electronic system.

When the test items are syllables, the score is referred to as "syllable articulation," with the terms "word articulation" and "sentence articulation" applying in the appropriate cases. Definite relationships have been found (1) (10) between scores for syllables, words, and sentences, which show that articulation generally increases with the length of test item. Hence scores for spondees are higher than for monosyllables, and scores for sentences are higher than for spondees.

The many applications of articulation testing might be described as a group by the broad term "intelligibility studies." Although the applications are quite varied (11,12,13,14,15), nearly all of them can be fitted into the basic structure of Figure 1. The interests of different groups of researchers generally center on only one or two "blocks" of this figure. Psychologists, for example, are more interested in the stimulus-response characteristics of the subjects, as opposed to the transmission properties of the channel, and hence are primarily concerned with the first and last blocks. For this group, the term "speech perception

test" has been suggested (16) as being more appropriate than "articulation test." Physicians and others specializing in speech audiometry are concerned with determining the optimum set of speech items with which to test the hearing acuity and discrimination loss of the listeners, the listeners in this case being patients with possible hearing defects (5,17,18,19,20, 21). Communications engineers may consider the subjects merely as "measuring devices" for determining the quality of the channel; in this application, as well as in the study of architectural acoustics, the center block of Figure 1 is of primary interest.

Whether one is testing the set of speech items, the channel, or the listeners, the procedure has become fairly well standardized (3,4). Although various parameters, such as the number and type of speech items, the number of listeners and their hearing acuity, instructions, and training, may be used as test variables, the most commonly-used variable is some parameter of the channel such as gain, frequency response, or amount of noise added to the speech signal. When the properties of a channel are varied by controlling a single variable such as signal/noise ratio, the articulation scores are commonly plotted as a function of that variable to yield an articulation curve, as shown in Figure 2 for words as test items. Such curves are also referred to as "gain functions" when the variable is channel gain (22). One exception to the use of articulation curves for presenting test data is the case where an information-theoretic presentation, such as a confusion matrix (23), is desired.

The word score, at a given point on the curve of Figure 2, is a function not only of the value of signal/noise ratio and other fixed factors such as the particular words and speaker used, but also of

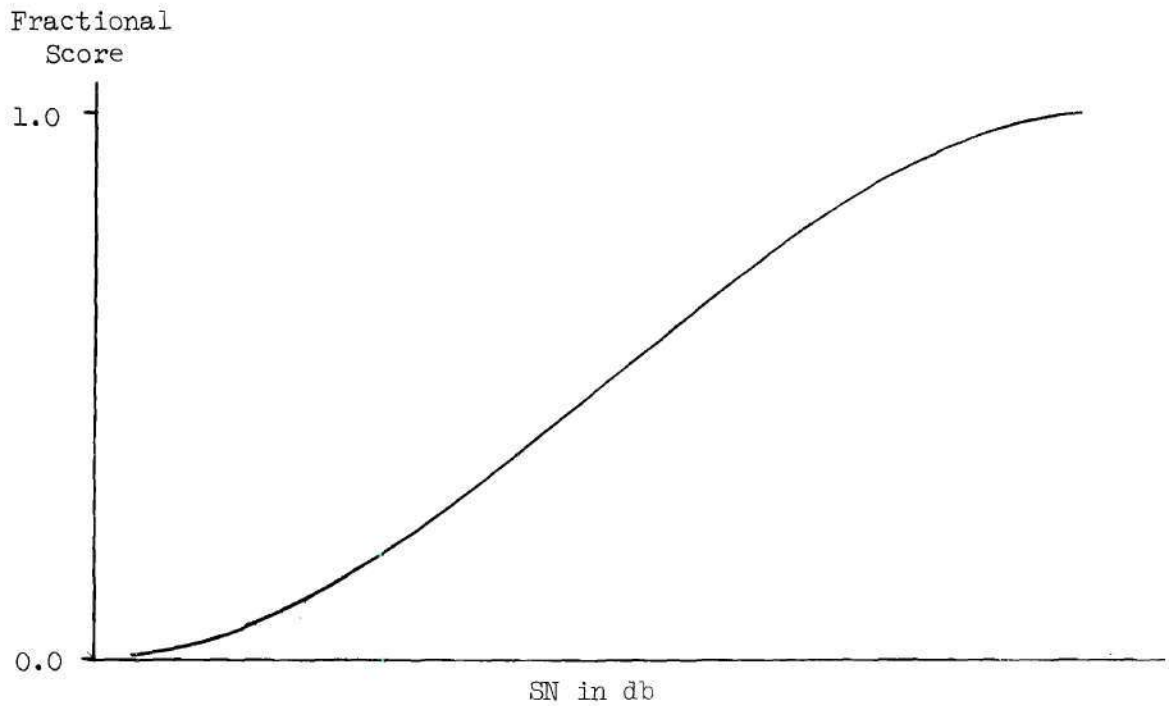


Figure 2. Articulation Curve for Words in Noise.

various subjective and physiological factors associated with the listeners, such as their auditory acuity, degree of training, and interest in the test. Such factors often change in a random way from day to day and test to test, implying that "word score" or "articulation curve" are not quantities capable of precise reproduction, even in tests where conditions are apparently identical (24, p. 389). Although word scores have the characteristics of a random variable, the reliability of the mean score (average over a number of tests or listeners) can be increased by increasing either the number of tests or the number of listeners (3, 25), and the terms "score" or "curve" are often used to denote such averages. If subjective factors can be minimized, as by careful selection and training

of listeners, standardization of test conditions, and the use of average scores, the resulting values of articulation constitute a fairly reliable measure of "speech intelligibility" at the channel output. Indeed, by its very definition, the "intelligibility" cannot be measured directly except with human listeners. Thus, regardless of the variation in detail or motivation involved in making articulation tests, such tests occupy an important place in intelligibility studies.

Aside from the inherent random nature of articulation scores, the results of different investigators are not always directly comparable, due to differences in speakers and listeners, in training and instruction of listeners, in definition and measurement of quantities such as signal/noise ratio, and in test conditions. Even the results of a single investigator using standardized procedures may not be comparable in some cases. For example, consider the difficulty of maintaining stable test conditions over long periods of time, as listeners experience a change in motivation or interest in the tests or even drop out of the team, necessitating replacements. Other changes may gradually occur in equipment, or in recordings of speech items, due to extensive use. A not insignificant effect is due to "learning" by the listeners (1,24,25,26). Although learning rate decreases with experience, there is evidence that learning never ceases entirely, and hence causes a relative increase in scores as time goes on.

To illustrate this problem, consider the case where articulation curves of noise-masked monosyllabic words are desired for a large number of word sets, and suppose further that many of the words are common to two or more sets. Intuitively, it is wasteful of testing time to test these

common words more than once, assuming that the influence of factors such as size of word sets and context (particular words in a set) is known. For definiteness, let the total number of different words be 1,000, and let each subset to be tested consist of exactly 50 words. The 1,000-word "master set" contains a fantastically large number (approximately 10^{87}) of distinct 50-word subsets, all of which are different by at least one word. Even if only 100 of these subsets are of interest, the number of word transmissions required to determine an articulation curve for each set is approximately 50,000. This large number of transmissions results partly from having to transmit a given word several times at each signal/noise ratio if the word is common to several subsets. A really exhaustive study, involving other possible 50-word subsets and subsets containing different numbers of words, might well be impractical from a standpoint of stability of test conditions over the extended period of testing time required, let alone the economic considerations involved. It would clearly be advantageous to have a procedure for predicting the various subset curves from tests on the master set only. Such a procedure, an example of which is developed in the following pages, would yield curves which were directly comparable, since they would have been obtained from the same basic set of data, the acquisition of which would have taken a relatively short time from the point of view of test condition stability.

Because of the tediousness of making articulation tests, much effort has been expended toward the development of estimation procedures which could predict articulation scores, particularly in the case of linear channels corrupted by masking noise (27). Most of these attempt to predict word scores from simple acoustical measurements, coupled with

tone thresholds for a human subject with normal hearing. The method involves determination of the frequency response of the channel, together with the spectral characteristics of the speech and noise, and is essentially limited to linear systems.

The best-known of these estimation procedures is the Articulation Index approach begun initially at Bell Labs (2,24,26,28,29,30), although other spectral approaches, such as that of Tkachenko (31), have been developed. At least four variations of the original Articulation Index method have been devised, mainly in an attempt to decrease the large number of tedious calculations required by the original method. The number of required calculations is one disadvantage of such procedures; another is the lack of accuracy and generality in many cases (10,32). Of significance is the fact that such methods do not take into account the particular words or other speech units to be transmitted. The results, while fairly accurate for "average" speech, can be in error for particular choices of words. The prediction method described in following pages avoids this problem by taking into account the contribution of individual words to the articulation curve.

CHAPTER II

FORMULATION OF PROBLEM IN TERMS OF THRESHOLDS

Nature of the Threshold Approach

Thresholds and Gain Functions

The threshold concept, as applied to the psycho-acoustical study of stimulus-response phenomena, is a relatively old one, one of the earliest determinations of auditory thresholds being made in 1876 (33). Various types of threshold have been defined and used in the study of hearing, the most common one being the threshold for pure tones. Tone threshold, also referred to as hearing acuity, threshold of detectability, or threshold of audibility, is the smallest effective sound pressure of pure tone which is just detectable by a human subject. It varies with the frequency of the tone and from subject to subject, being about 0.0001 dynes/cm^2 under optimum conditions (16). Thresholds have also been determined for interrupted tones, frequency-modulated tones, and tones masked by an interfering sound such as other tones or noise. Both masked thresholds and thresholds with no masking have been determined for other acoustic stimuli, notably clicks, noise, and speech.

The somewhat crude concept of a threshold as a level above which a stimulus is always perceptible, and below which it is never perceptible, is not strictly applicable to any situation involving human subjects. Because of the subjective randomness involved in the responses, a given stimulus level is sometimes detected, and sometimes not, over some range of levels which might be called a "region of ambiguity" or "threshold

region." Thus a better definition of threshold is that level which is detected a certain percentage of the time, during a number of trials (24). One might then define many types of threshold, i.e., "50 per cent threshold," "90 per cent threshold," or "100 per cent threshold." The last example naturally raises the question as to whether there is a minimum level above which the stimulus is always detected, or, conversely, a maximum level below which the stimulus is never detected, such points logically serving to define the threshold region. Aside from the purely practical consideration of not being able to experimentally determine such levels because of the words "always" and "never," the answer is intuitively "no." Furthermore, the experimentally determined percentage of detections at any given level will be found to fluctuate with the number of trials made, although this percentage may tend toward some fixed number (the probability of detection) as the number of trials increases without limit. Instances in which, for all practical purposes, a single "all-or-none" threshold may be specified are the cases where the width of the threshold region is considerably less than the experimental error in setting the level, or where the increments in level are required to be larger than the width of the threshold region. Not only is the nature of this region indeterminate in such cases, but the very fact of its non-zero width is insignificant for experimental purposes.

As applied to speech, several types of threshold have been used, in addition to threshold of detection. The thresholds of perceptibility and of intelligibility have been defined (3) as the levels at which the listener can, respectively, " ...understand, with difficulty, the gist of connected discourse.," and " ...just ...obtain without perceptible effort

the meaning of almost every phrase of connected discourse ..." When particular speech units are used as stimulus items, the resulting thresholds can be referred to as "word threshold," "vowel threshold," "phoneme threshold," etc., and these can be further particularized to include masking with various sounds. The word "threshold," then, when applied to speech without qualification, is somewhat vague; it has precise meaning only when a quantitative definition is given and the stimulus item specified.

The definition of noise-masked word threshold used here is that value of signal/noise ratio at which the word is correctly identified in 50 per cent of the trials, i.e., the value at which the probability of correct reception is 0.5. The use of 50 per cent thresholds in studies of the non-masked case is common, this definition having been utilized for words (5,20), vowels (34), and alphabet letters (22).

Clearly, the level at which other percentages of identification occur can be found, in the non-masked case, by varying the gain of the test system; such data, as noted earlier, are often plotted in the form of gain functions. By extension, the term "noise-masked gain function" (NMGF) is used here to denote the articulation curve for a single word imbedded in noise, plotted as a function of signal/noise ratio, while the term "articulation curve" is reserved for the curve pertaining to a set of such words. Word threshold is then the abscissa of the 50 per cent point on the NMGF, expressed in db of signal/noise ratio. The use of the ratio of speech to noise level as the independent variable is conventional, primarily because the scores are not a function of speech level or of noise level alone, but of their ratio, at least above a certain

minimum noise level. That thresholds depend on this ratio alone, above some minimum noise level, has been shown by Hawkins and Stevens (15), although exceptions to this were later found by Kryter (11) at extremely high speech and noise levels. In the studies described here, the use of signal/noise ratio as the test variable permits the measurement of speech and noise powers at any point in the linear test system, and permits acoustical measurements on sound waves to be replaced by electrical measurements on the corresponding voltage waveforms.

Based on the supposition that the intelligibility characteristics of words are adequately described by their NMGF's, the nature and describing parameters of such functions become of interest, as does the question of whether these describing parameters alone constitute a valid description of the "intelligibility" of words. Non-masked gain functions have been found to have roughly the same general shape as the curve in Fig. 1, and parameters which have been used to describe such functions for speech units include 50 per cent threshold and slope in the vicinity of the 50 per cent point. Essentially these same parameters were found useful in the present study for the masked case, except that "spread" of the NMGF was used instead of slope. The spread, or width of threshold region, is approximately inversely proportional to the slope, and hence contains the same information. Although much work has been done on the selection of word sets having similar articulation curves, or having curves which are very steep, relatively little has been done to formally relate the articulation curve to the gain-function parameters of the individual words in the set.

The Effect of Intensity Differences

That the relative intensity or power of speech units is a factor in recognition, has been known for some time. Indeed, this factor is the principle one in determining relative intelligibilities in the case where all of the speech units have essentially the same threshold, i.e., where the threshold spread is much less than the intensity spread, and this fact has at times obscured the effects of threshold differences between words. In 1947 Hudgins and others (18) described two methods for obtaining homogeneity in a set of speech items (as evidenced by the steepness of the gain function), based on the assumption that differences in basic audibility are due to intensity differences alone. Harris, in a paper (20) published two years later, pointed out that equal intensity does not guarantee equal intelligibility. Using recorded versions of selected PB words, he showed that the spread in thresholds was not significantly changed by processing words for equal intensity, but was reduced from 42 db to 26 db by equating words for threshold. The resulting gain function for the word set had a slope approaching that of spondees in steepness. In 1950 Curry (34), concluded that " ...some factor in addition to intensity is contributing to the identification ...," after making tests on nine American vowel sounds to determine their relative intensities and thresholds. When he calculated relative thresholds, corrected for intensity, a spread of 2.75 db still remained, causing him to state that " ...some factor, unique to each vowel, in addition to intensity makes the identification of each vowel possible." In retrospect, it seems apparent that this factor is the difference in thresholds among vowels. It should be noted that the threshold region for vowels was apparently too narrow, compared to the

experimentally used increments in gain, to permit the recognition or specification of individual vowel gain functions. Hirsh and others (5) in 1952 confirmed Hudgins' results by equalizing intensities of spondaic words, resulting in CID Test W-1. Due, perhaps, to test procedures, no significant threshold differences between spondees was discovered, although the lists were found to be not equal in difficulty, indicating that some (small) spread in thresholds did exist.

Finally, Curry and others reported (22) in 1960 their investigation of the intelligibility of non-masked alphabet letters. Points on the gain functions for individual letters were obtained, using a 5 db increment in gain, and these were visually fitted with smooth curves. These, in turn, yielded 50 per cent thresholds and also slopes in the 20-to-80 per cent region. Large differences were found in the shape and location of curves, with a spread in thresholds of 17.75 db, part of which was due to intensity differences. When the data was adjusted so as to equate thresholds, the pooled curve (averaged over all speakers, listeners, and letters) exhibited a steeper slope than before adjustment. Upon adjustment for intensity differences, the spread in thresholds exhibited only a one db change, confirming earlier results. Finally, adjustment of the data for intensity differences resulted in a pooled curve with somewhat higher slope, indicating that at least part of the spread in this curve is due to intensity differences.

To summarize, individual speech elements with no masking characteristically exhibit a spread in both intensity and threshold, both of which contribute in some way to the spread and steepness of the articulation curve for a set of such elements. The present study extends these concepts

to the noise-masked case, and in the following section, shows how the parameters of word power, noise power, and NMGF threshold and spread are mathematically related to the articulation curve.

Mathematical Formulation

General

The prediction of articulation curves, in the sense used here, stems from the intuitive idea that if one knows the signal/noise threshold of each of a set of words, and knows that the set is to be transmitted at a given signal/noise ratio, one ought to be able to predict which words will be understood, i.e., which words are received above threshold. The fraction of words above threshold is then the articulation score. Such a scheme, depending on a word being either "above threshold" or "below threshold" is clearly valid only when the NMGF's are step-functions, i.e., when a single "all-or-none" threshold is adequate to describe the situation. A high degree of validity may still remain in the case of non-zero NMGF spread, provided the thresholds are sharp; in this case the NMGF spreads will be small and a single "all-or-none" threshold might be specified at the 50 per cent level. In the case where NMGF's have large spreads, the threshold alone may still suffice for use in a prediction scheme, provided the number of words is large enough. As an alternative to replacing the NMGF with a step function, the actual spread may be taken into account by use of a linear approximation in the threshold region. Both of these approaches are considered in the following.

In addition, the problem of defining signal/noise ratio for individual words and sets of words is considered, so that the articulation curve can be predicted without a tedious inspection of each word at each level.

These and other quantities are defined as introduced, and a summary of all such definitions is included in Appendix C. In general, capital letters are used for quantities expressed in db and the corresponding lower case letter is used when the quantity is not in db, while most Greek letters represent quantities expressed in db.

Definition of Signal/Noise Ratio

Before calculating, for an individual word, the ratio of word power to noise power, the latter quantities must be precisely defined. Consider a set of words recorded at intervals on one track of a magnetic tape, an adjacent track on the same tape containing recorded noise. When the tape is played back, the speech and noise signals appear as electrical waveforms which, when linearly added, produce a set of noise-masked words in the form of an electrical signal. When the relative levels of speech and noise are varied by means of an attenuator in the speech channel, it is assumed that all words are attenuated equally, thus preserving their relative powers and at the same time altering the mean power of the set. The power of a given word is defined as the average electrical power in the word waveform, the average being taken over the time duration T of the word. It will be assumed that this power is to be measured at a specific point in the playback system and for a standard setting of the playback gain control, so that a well-defined and reproducible voltage waveform e(t), of duration T, exists for a given word. If this waveform is produced across a resistive impedance R at the point of measurement, the word power p is the average power dissipated in R by e(t); this quantity is given, for the i^{th} word, by

$$p^i = \frac{1}{T^i} \int_{T^i} \frac{\left[e^i(t)\right]^2}{R}\, dt = \frac{w^i}{T^i}, \qquad (1)$$

where w^i is the word energy. In the case of N words in the set ($i = 1,2,3,\ldots,N$), the mean word power over the set is given by

$$\bar{p} = \frac{1}{N} \sum_{i=1}^{N} p^i, \qquad (2)$$

and the word power in db relative to the mean power is

$$P^i = 10 \log \frac{p^i}{\bar{p}}. \qquad (3)$$

Each word is masked by the section of noise waveform lying adjacent to it on the magnetic tape; this part of the noise is always played back in time coincidence with the word. Thus a "noise word" of duration T and voltage $e_n(t)$ is defined, having power

$$p_n^i = \frac{1}{T^i} \int_{T^i} \frac{\left[e_n^i(t)\right]^2}{R}\, dt = \frac{w_n^i}{T^i} \qquad (4)$$

for the noise associated with the i^{th} word, provided the noise voltage is also developed across R. The mean noise power, over the set of "noise words," is

$$\bar{p}_n = \frac{1}{N} \sum_{i=1}^{N} p_n^i, \qquad (5)$$

and the noise power in db relative to the mean noise power is

$$P_n^i = 10 \log \frac{p_n^i}{\bar{p}_n} \quad . \tag{6}$$

As it was for the speech, it is assumed that varying the noise level by means of an attenuator in the noise channel will cause all noise words to be equally attenuated, thus preserving their relative power while at the same time altering the mean noise power. The extension of this idea to the case of gain inserted in either channel, or of gain in one channel and attenuation in the other, is apparent.

It is now possible to precisely define the signal/noise ratio for the i^{th} word, as follows:

$$sn^i = \frac{p^i}{p_n^i} \quad , \tag{7}$$

or, in db,

$$SN^i = 10 \log sn^i = P^i - P_n^i + 10 \log \frac{\bar{p}}{\bar{p}_n} \quad . \tag{8}$$

The last quantity can logically be defined as the signal/noise ratio SN for the <u>set</u> of noise masked words.

$$SN = 10 \log \frac{\bar{p}}{\bar{p}_n} = 10 \log sn \quad , \tag{9}$$

where sn is the mean word power for the set divided by the mean (masking) noise power for the set. Other possible definitions for sn, such as the

mean value of sn^i or $\Sigma w^i/\Sigma w_n^i$ have no apparent advantage over the one given above; furthermore, the definition given in equation (9) is somewhat more easily incorporated into a mathematical formulation of the prediction scheme. As a final justification, the experimental values obtained with these various definitions are later shown to be almost identical for $N \geq 20$.

The group signal/noise ratio SN can be varied by changing $\bar{p}$, $\bar{p}_n$, or both. Note that when this is done, P^i and P_n^i do not change, since they are measures of relative power, with respect to the mean. SN is the test variable for obtaining the articulation curve of a word-set, and is measured along the abscissa of such a curve. It is, in general, different for different word-sets, hence the P^i depend upon the word-set involved.

Formulation for Step-Like NMGF's

Consider the case of a word having a very sharp intelligibility threshold. As SN for the set is increased, the SN^i also increase, by the same amount. When the SN^i of the i^{th} word reaches the threshold value, denoted by α^i, the word suddenly emerges from the noise and becomes intelligible, resulting in the step-like NMGF of Figure 3. The variable α^i, then, is the smallest value of SN^i for which the word is intelligible, and hence

$$SN^i \geq \alpha^i \Longrightarrow i^{th} \text{ word is understood ,} \tag{10}$$

and

$$SN^i < \alpha \Longrightarrow i^{th} \text{ word is not understood.} \tag{11}$$

In actuality, no such threshold behaviour ever occurs, and hence the step

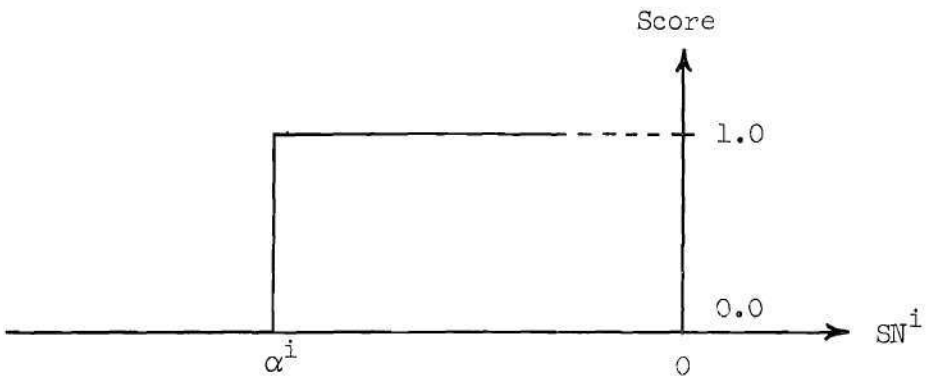


Figure 3. Step-Like NMGF.

of Figure 3 may be considered as simply an approximation to the actual NMGF in its very narrow threshold region.

An "intelligibility variable" ξ may now be defined for the i^{th} word by

$$\xi^i = \alpha^i - SN^i , \tag{12}$$

so that

$$\xi^i \leq 0 \Longrightarrow i^{th} \text{ word is understood,} \tag{13}$$

and

$$\xi^i > 0 \Longrightarrow i^{th} \text{ word is not understood.} \tag{14}$$

Using previous definitions, ξ^i can now be put into a more useful form, as follows:

$$\xi^i = \alpha^i - SN^i = \alpha^i - 10 \log \left[\frac{p^i}{p_n^i} \frac{\bar{p}}{\bar{p}} \frac{\bar{p}_n}{\bar{p}_n} \right] \tag{15}$$

$$= \alpha^i - 10 \log \left[\frac{p^i}{\bar{p}} \frac{\bar{p}_n}{p_n^i} \frac{\bar{p}}{\bar{p}_n} \right]$$

$$= \alpha^i - 10 \log \frac{p^i}{\bar{p}} + 10 \log \frac{p_n^i}{\bar{p}_n} - 10 \log \frac{\bar{p}}{\bar{p}_n}$$

$$= \alpha^i - P^i + P_n^i - SN \ .$$

In the last form, ξ^i is a function of SN only through the last term, since α^i, P^i, and P_n^i do not, by previous definitions and assumptions, vary with either $\bar{p}$ or $\bar{p}_n$ for a given word-set. Hence the first three components of ξ^i may be combined into a new variable β^i which is independent of SN, yielding

$$\xi^i = \beta^i - SN \ , \tag{16}$$

where

$$\beta^i = \alpha^i - P^i + P_n^i \ . \tag{17}$$

The quantity β^i is seen to be an "adjusted" threshold, where α^i has been adjusted for word power and noise power variations among the words. Thus β^i is the value of SN at which the word becomes intelligible, and thus β^i varies from one word-set to another having a different SN, while α^i does not so vary.

It will now be shown that the articulation score is $F_\beta(SN)$, i.e., the distribution function of β evaluated at the desired signal-noise ratio.

From the N experimentally determined values of ξ^i the cumulative distribution function F_ξ can be plotted, as shown in Figure 4. By definition, this function, evaluated at x, is the fraction of ξ's having values equal to or less than x, and hence for finite values of ξ is a monotone curve which increases from zero at the smallest value of ξ to unity at the largest value. For finite N, the curve is a series of steps; for a continuous distribution of ξ's it is a smooth curve as shown.

The ordinate of F_ξ at x_1 is a probability, namely, the probability that a ξ chosen at random, i.e., for a randomly chosen word, will be equal to or less than x_1, assuming that all words are equally likely choices. Thus $F_\xi(0)$ is the probability that a randomly-chosen word will have a ξ which is equal to or less than zero, i.e.,

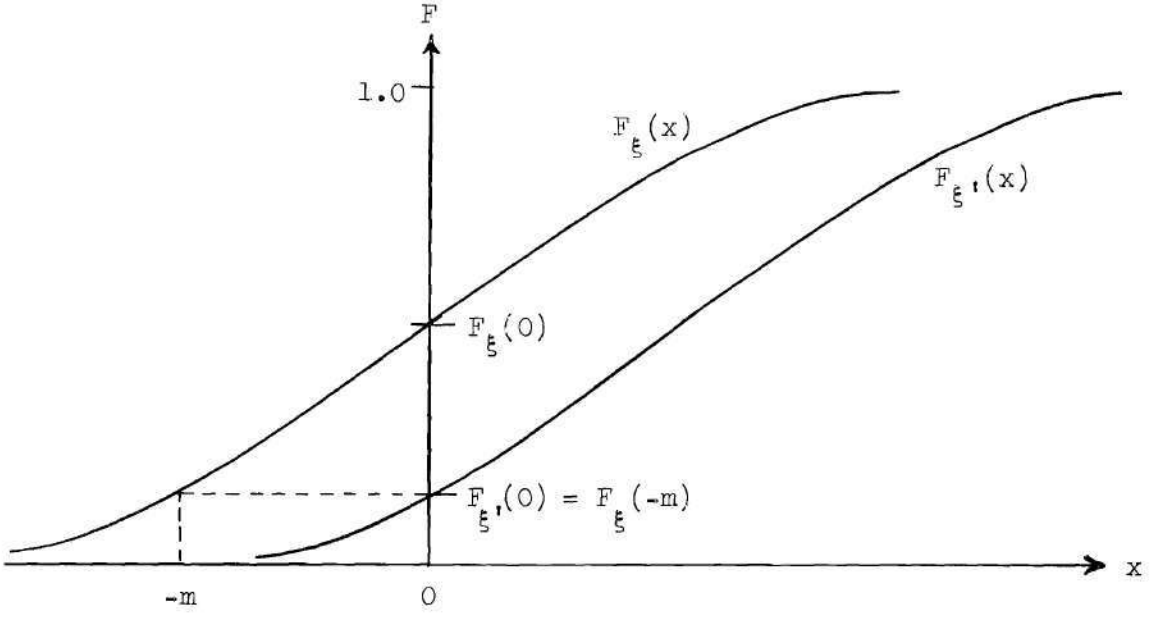


Figure 4. Distribution Functions for ξ and ξ'.

$$F_{\xi}(0) = \Pr\left[\xi^{i} \leq 0\right], \tag{18}$$

where superscript "i" has been used to identify the randomly-chosen word. The event in brackets of equation (18) is identical with the event, "the randomly-chosen word is understood," from equation (13), and the probability of this event is equal to the fraction of words understood when all words in the set are transmitted (26, p. 280). Hence the fractional articulation score is given by

$$\text{Fractional score} = \Pr\left[\text{a randomly-chosen word is understood}\right] \tag{19}$$

$$= F_{\xi}(0) .$$

The fractional score may be obtained, then, by reading $F_{\xi}(0)$ from the plotted curve, as shown in Figure 4; this number, multiplied by 100, is the per cent word score.

The values of ξ, and hence the location of F_{ξ} (but not its shape), vary with SN. A possible value of SN to use in calculating ξ's and plotting the curve is the value at the standard measurement point, but a more useful result is obtained by plotting for SN = 0. For this choice, ξ is identical with β, from equation (16), and hence F_{ξ} is identical with F_{β} and one can write

$$F_{\xi}(x) = F_{\beta}(x) , \qquad SN = 0 . \tag{20}$$

Assuming this is done, consider the effect on F_{ξ} of transmitting a word set at some non-zero value of SN, say at SN = -m, where $m > 0$. Each of the SN^{i}'s will be decreased by m db from its value for SN = 0 and hence,

from equation (12), each of the ξ's will increase by m db. The resulting new values of ξ, denoted by ξ', have a new distribution function $F_{\xi'}$, where

$$\xi' = \xi + m . \tag{21}$$

It follows that

$$F_{\xi'}(x) = F_{\xi}(x - m) , \tag{22}$$

and hence

$$\text{Fractional score} = F_{\xi'}(0) = F_{\xi}(-m) \tag{23}$$

The curve plotted for SN = -m is the original curve (SN = 0) shifted to the right by m db, and the fractional score may be read either from the new curve at the origin or from the original curve at -m, as shown in Figure 4. This latter is equivalent, by equation (20), to evaluating F_{β} at -m, and hence

$$\text{Fractional score} = F_{\beta}(-m). \tag{24}$$

By similar reasoning, if the words are transmitted at SN = m, then

$$\text{Fractional score} = F_{\beta}(m). \tag{25}$$

Replacing the specific values -m or m by the general variable SN permits the score to be expressed very compactly by

$$\text{Fractional score} = F_{\beta}(SN). \tag{26}$$

To summarize for the sharp-threshold case: if the thresholds α are obtained, along with values of individual word power p and noise power p_n, a value of β can be calculated for each word, these values being independent, for a fixed word-set, of SN at the point where p and p_n are measured. F_β can now be plotted, and the fractional word score read directly from this curve for any value of SN. Alternatively, if F_β can be approximated by a closed-form mathematical expression, scores can be calculated directly from this expression, using equation (26).

Formulation for Broad Threshold Regions

In the case where thresholds are not sharp, the NMGF (or some approximation to it) would logically serve better than a single "threshold" to define a given word's contribution to the articulation curve. Clearly, the set of actual NMGF's define the articulation curve exactly, since, as described in Chapter III, they constitute simply an alternative way of presenting the same data as is used to plot this curve. It is desirable to represent the collection of NMGF's by a reduced set of data, such as a collection of describing parameters, in which each NMGF is represented by numbers characterizing its shape and location.

Assuming that the NMGF's vary in a fairly smooth and monotone fashion from zero to unity fractional score, one might obtain good representations of these curves by means of piece-wise linear approximations, as illustrated in Figure 5. Such an approximation, referred to henceforth as a linear approximation, is completely characterized by two parameters: α^i, arbitrarily defined as the value of SN^i for 0.5 fractional score and referred to as the "threshold"; and Δ^i, the total spread in db of the threshold region defined by the approximating curve. Not only do α^i and

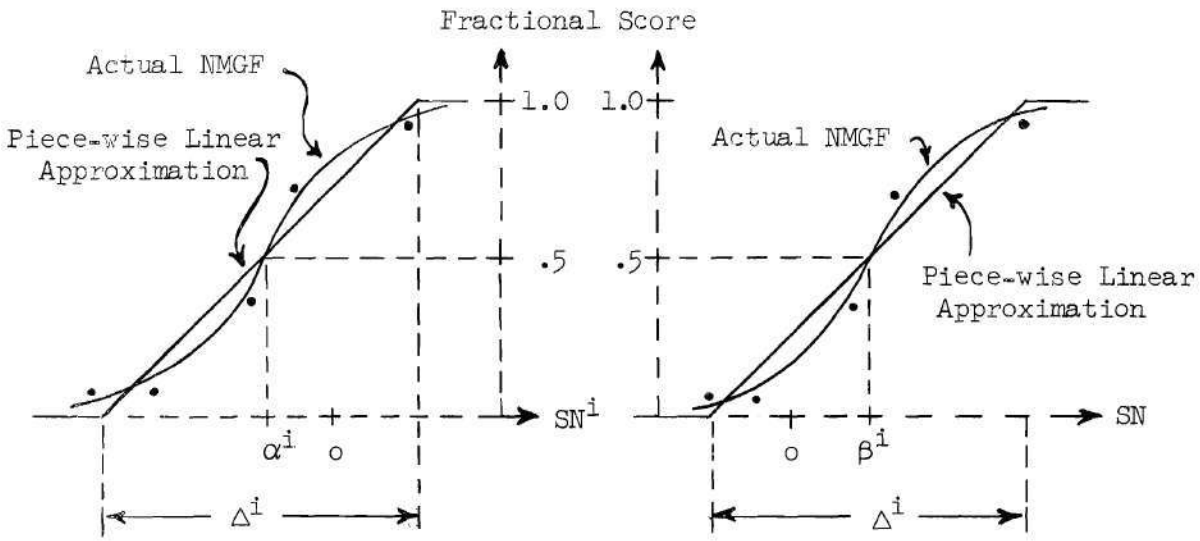


Figure 5. NMGF and Piecewise Linear Approximations.

Δ^i serve to define the linear approximation, but also, to the extent that the approximation is a close one, they convey information about the actual spread and threshold of the NMGF. In the latter sense, α^i and Δ^i can be considered as parameters of the i^{th} word which describe its intelligibility properties.

The linear approximation can be expressed mathematically as a function of either word signal/noise ratio SN^i or, for a given word-set, as a function of group signal-noise ratio SN, the relationship between these variables being, from equations (3), (6), (8), and (9),

$$SN = 10 \log \frac{\bar{p}}{\bar{p}_n} = 10 \log \left[\frac{p^i}{p^i_n} \frac{\bar{p}}{\bar{p}^i} \frac{p^i_n}{\bar{p}_n} \right] \qquad (27)$$

$$= SN^i - P^i + P^i_n \, .$$

In the case where NMGF's of arbitrary words are to be compared, the more basic variable SN^i may be used; in the case where a set of NMGF's is to be studied for its relation to the articulation curve, SN is more appropriate as the variable. In the latter case, the threshold value is β^i and the linear NMGF approximation may be expressed mathematically as

$$G^i(SN) = \begin{cases} 0, & SN < \beta^i - \frac{\Delta^i}{2} \\ \frac{1}{\Delta^i}\left[SN + \frac{\Delta^i}{2} - \beta^i \right], & \beta^i - \frac{\Delta^i}{2} \leq SN \leq \beta^i + \frac{\Delta^i}{2} \\ 1.0, & SN > \beta^i + \frac{\Delta^i}{2} \end{cases} \quad (28)$$

If each of N words is assumed to contribute independently to the articulation curve, then intuitively the curve is expressible as the sum of N individual NMGF's, each one being weighted by a factor of $\frac{1}{N}$. This can be justified mathematically by interpreting a point on $G^i(SN)$ as the probability that the i^{th} word will be understood when the word set is transmitted at SN. Then,

$$\text{Fractional score} = \text{Prob}\left[\text{a randomly chosen word is understood} \right] \quad (29)$$

$$= \text{Prob}\Big[(\text{the first word is chosen } \underline{\text{and}} \text{ the first word is understood}) \underline{\text{or}} (\text{the second word is chosen } \underline{\text{and}} \text{ the second word is understood}) \underline{\text{or}} \ldots . (\text{the } N^{th} \text{ word is chosen } \underline{\text{and}} \text{ the } N^{th} \text{ word is understood}) \Big]$$

$$= \sum_{i=1}^{N} \text{Prob}\left[i^{th} \text{ word is chosen } \underline{\text{and}} \ i^{th} \text{ word is understood} \right]$$

$$= \sum_{i=1}^{N} \text{Prob}\left[i^{th} \text{ word is chosen} \right] \text{Prob}\left[i^{th} \text{ word is understood} \right]$$

$$= \sum_{i=1}^{N} \frac{1}{N} \, G^{i}(SN) = \frac{1}{N} \sum_{i=1}^{N} G^{i}(SN) \, ,$$

since each of the words is an equally likely choice. Once the G^i are determined in the form given by equation (28), the articulation curve is completely described by equation (29). The accuracy of this description depends on how closely the G^i functions approximate the actual NMGF's. A practical way of obtaining as many points as desired on the articulation curve, assuming the Δ^i and α^i are known, would be to use a digital computer programmed to compute equation (29). If a relatively small number of words are involved, and if their G^i functions are available, points can be read or computed by hand and summed to give the word score.

Modifications for Various Cases

If, as was found to sometimes be the case (see Chapter III), the NMGF points in the signal/noise region of interest are best fitted by a negative-slope line, then certain modifications must be made in the mathematical expressions derived above. Such a situation might arise if the test range (of SN) of interest does not cover the entire threshold region of one of the NMGF's, and if in addition the NMGF is not monotone increasing but has a region of negative slope which falls in the test range. Two examples are shown in Figure 6.

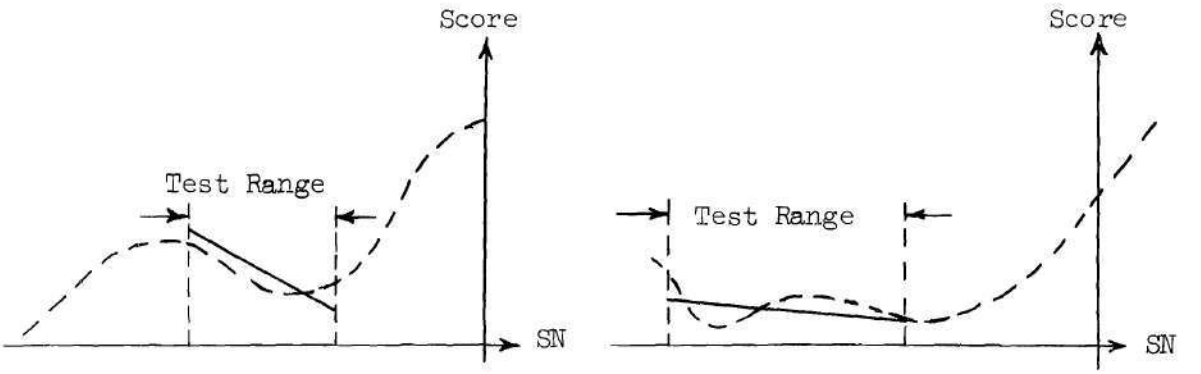


Figure 6. NMGF's with Negative Slopes in the Test Range.

The situation illustrated may be aggravated if only a small number of points is available for plotting the NMGF and/or if there is considerable possible error in any given point. At any rate, the fact that the slope of the linear approximation is negative does not affect the accuracy of the procedure described above, provided a good fit to the curve is obtained in the test region and provided the procedure is applied only there. It is true that the linear approximation no longer represents the NMGF as a whole, and hence values of α^i and Δ^i determined from such lines are not valid descriptors of the NMGF threshold and spread.

For words exhibiting negative-slope G^i functions, the definition given by equation (28) is replaced, in the test region, by

$$G^i(SN) = -\frac{1}{\Delta^i}\left[SN - \frac{\Delta^i}{2} - \beta^i\right], \qquad (30)$$

equation (29) then being used as before to plot the articulation curve.

In using the method based on a step-approximation to the NMGF, the negative-slope G^i functions present somewhat more difficulty. For calculations in the test range, this method is still valid provided the α^i obtained from the G^i function is interpreted as shown in Figure 7. Here, α^i, the 50 per cent point on G^i, is interpreted as a "reverse threshold," i.e., a value of SN^i below which the word is understood and above which it is not understood. Such an approximation can never be valid either for representing the entire NMGF or for describing the intelligibility properties of the word, but it may still be employed as a mathematical substitute for the NMGF in the test region, and when so used is as valid as the more typical step of Figure 3. The procedure may be developed as follows.

Consider a set of $n_1 + n_2 = N$ words, where n_1 words have positive-

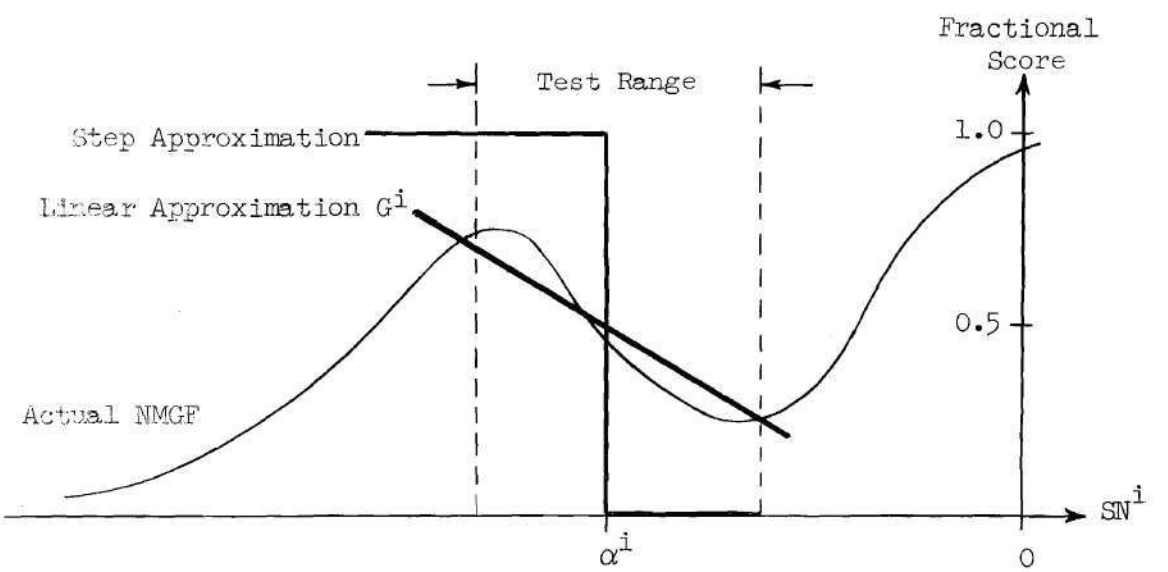


Figure 7. NMGF with Negative-Slope Linear Approximation and Reversed Step Approximation in Test Range.

slope G^i functions and n_2 words have negative-slope G^i functions. If p^i and p_n^i are measured for each of the words, then SN for the N-word set is determined. Values of ξ^i and β^i are calculated from equations (12) and (17) for the n_1-word set and denoted by ξ_1^i and β_1^i, respectively. Similarly, values ξ_2^i and β_2^i are obtained for the n_2-word set. The fractional word score is given by

$$\begin{aligned}
\text{Fractional score} &= \Pr\left[\text{a randomly-chosen word is understood}\right] \qquad (31)\\
&= \Pr\left[\text{the word is in } n_1\right]\Pr\left[\text{an } n_1 \text{ word is understood}\right]\\
&+ \Pr\left[\text{the word is in } n_2\right]\Pr\left[\text{an } n_2 \text{ word is understood}\right]\\
&= \frac{n_1}{n_1+n_2}\Pr\left[\text{an } n_1 \text{ word is understood}\right]\\
&+ \frac{n_2}{n_1+n_2}\Pr\left[\text{an } n_2 \text{ word is understood}\right]\\
&= K_1 Pr_1 + K_2 Pr_2 \,.
\end{aligned}$$

Now

$$\begin{aligned}
Pr_1 &= \text{Prob}\left[\xi_1 \le 0\right] = \text{Prob}\left[\beta_1 - SN \le 0\right] \qquad (32)\\
&= \text{Prob}\left[\beta_1 \le SN\right] = F_{\beta_1}(SN)\,,
\end{aligned}$$

and

$$\begin{aligned}
Pr_2 &= \text{Prob}\left[\text{an } n_2 \text{ word is } \underline{\text{below}} \text{ threshold}\right] \qquad (33)\\
&= \text{Prob}\left[\xi_2 > 0\right] = 1 - \text{Prob}\left[\xi_2 \le 0\right] = 1 - \text{Prob}\left[\beta_2 \le SN\right]\\
&= 1 - F_{\beta_2}(SN)\,.
\end{aligned}$$

Hence

$$\text{Fractional word score} = \frac{n_1}{n_1+n_2} F_{\beta_1}(SN) + \frac{n_2}{n_1+n_2}\left[1 - F_{\beta_2}(SN)\right]. \quad (34)$$

The articulation curve may now be plotted as described earlier, but using equation (34) in place of equation (26).

For the case where the noise power p_n^i is constant from word to word,

$$p_n^i = \bar{p}_n , \qquad i = 1,2,\ldots,N . \quad (35)$$

Such a case might arise, for example, when relatively long-duration words, such as spondees, are masked by band-limited white noise. The calculation of certain quantities is then made easier, since

$$P_n^i = 10 \log \frac{p_n^i}{\bar{p}_n} = 0 \quad \text{and} \quad (36)$$

$$\beta^i = \alpha - P^i + P_n^i = \alpha^i - P^i . \quad (37)$$

In computing the "corrected threshold" β^i (threshold value of SN), equation (37) reveals that α^i need be corrected only for variations of word power, i.e., there is no longer any contribution to the spread of the articulation curve from a spread in noise power.

CHAPTER III

EXPERIMENTAL INVESTIGATION OF NOISE-MASKED GAIN FUNCTIONS

General Approach

It has been shown in Chapter II that the articulation curve is related in a specific way to the word parameters p, p_n, α, and Δ. The experimental measurement of these and related parameters is now described, for the set of 40 monosyllabic test words in Table 1 below.

Table 1. Master Set of Test Words

ache	deck	jam	please	thrash
bald	dill	law	pulse	toil
bead	fame	leave	rate	turf
cast	fig	lush	rouse	vow
check	flush	muck	shout	wedge
class	gnaw	neck	size	wharf
crave	gob	nest	stag	who
crime	hurl	path	take	why

The test words were selected from lists 3 and 20 of the Harvard phonetically-balanced word lists. The number of words was chosen as being large enough to indicate the range of variation in word parameters to be expected for such monosyllables and also large enough to permit a convenient decomposition into several subsets possessing articulation curves of different shapes. At the same time, the number was small enough

to permit the required measurements and calculations to be done in a reasonable amount of time.

The general approach taken in this experimental part of the research was to record the words and noise on magnetic tape, thus fixing the test material in a standard reproducible form, and to then measure, with instrumentation designed for this purpose, the word power, word duration, and noise power on these "master" tapes. The noise-masked words were then presented to a listening team at various signal/noise ratios, and the resulting individual word scores were processed to yield NMGF's for individual words and listeners. Finally, the NMGF parameters α and Δ were calculated, followed by calculation of quantities such as β.

The results, especially of the articulation tests, are relative, rather than absolute, since they involve specific choices of talker, listeners, and test environment. As many of the test variables as possible were fixed, and the tests arranged so that effects of unknown factors were as random as possible, thus clearly placing in evidence the effects of changing the controlled variables.

Power Measurements and SN Calculations

Various specific types of power have been defined (10,26) with respect to speech; the definition found most useful here is that given by equation (1) of Chapter II, in terms of the instantaneous speech voltage and impedance level. Except for instantaneous power, all definitions of speech power (as well as all devices for measuring it) involve an averaging interval; in the present study where this is taken to be the word duration, a means of measuring this duration was necessary. Essentially, the measurement process, based on equation (1), was split into two

operations, although these were performed simultaneously. A device ("word timer") was constructed to measure the duration T during which the word voltage waveform, at a standard measurement point, exceeded some pre-set threshold voltage. A second device ("energy meter") was constructed to measure, at the same point, the energy w in the waveform, where, from equation (1),

$$w^i = \frac{1}{R} \int_{T^i} \left[e^i(t)\right]^2 dt \tag{38}$$

for the i^{th} word, and where $p^i = w^i/T^i$.

Since both the impedance level R and the location of the standard measurement point involve arbitrary choices, the absolute values of w (and hence of p) are of less interest than relative values, particularly in view of the fact that only ratios of energy are used as basic variables in the equations. For example,

$$P^i = 10 \log \frac{p^i}{\bar{p}} = 10 \log \frac{w^i/T^i}{\frac{1}{N}\sum_{i=1}^{N} w^i/T^i}, \tag{39}$$

$$P_n^i = 10 \log \frac{p_n^i}{\bar{p}_n} = 10 \log \frac{w_n^i/T_i}{\frac{1}{N}\sum_{i=1}^{N} w_n^i/T^i}, \tag{40}$$

$$SN^i = 10 \log \frac{p^i}{p_n^i} = 10 \log \frac{w^i/T^i}{w_n^i/T^i} = 10 \log \frac{w^i}{w_n^i}, \tag{41}$$

and

$$SN = 10 \log \frac{\bar{p}}{\bar{p}_n} = 10 \log \frac{\frac{1}{N} \sum_{i=1}^{N} w^i/T^i}{\frac{1}{N} \sum_{i=1}^{N} w_n^i/T^i} \quad . \qquad (42)$$

It is clear from the above equations that the values of w may all be multiplied by a constant without changing the values of the quantities on the left; the same is true of the values of T. Another way of saying the same thing is that the units in which w, T, and p are expressed are immaterial. In all following curves and tabulations involving numerical values of these quantities, the units are non-standard, although they may readily be converted to mks units of joules, seconds, and watts as described below.

Because of the gain constants and read-out calibration of the timer and energy meter, the energy in joules and time duration in seconds of a word were not numerically equal to the values w and T read at the output, but were proportional to them through calibration constants K_1 and K_2 respectively, so that

$$\text{Word energy in joules} = (K_1)(w) \qquad (43)$$

$$\text{Time duration in seconds} = (K_2)(T) \qquad (44)$$

Since, for present purposes, the values of K_1 and K_2 are immaterial, the output readings w and T, as well as their ratio p, are used directly as numerical values for word energy, time duration, and word

power. As a matter of interest, K_1 was evaluated and found to be given by

$$K_1 = 1.63 \times 10^{-5} , \tag{45}$$

while K_2, from the nature of the timer operation, was known to be

$$K_2 = 10^{-3} . \tag{46}$$

Thus w and T, wherever numerical values are given, are in units of K_1 joules and milliseconds, respectively, and corresponding values of p (or of noise power p_n) are in units of 16.3 milliwatts.

A block diagram of the instrumentation is shown in Figure 8, with various points identified for relation to the voltage waveforms of Figure 10. All circuits except the squarer, 10 Kc oscillator, and indicating instruments were built up on a Donner Scientific model 3100 Analog Computer, utilizing the computer's dc operational amplifiers, precision potentiometers, and standard reference voltages. Additional external components such as resistors, capacitors, and fast solid-state diodes, were utilized by plugging them into the computer's problem board, while connections to the externally-mounted relay and to various floating voltage sources (dry cell batteries) were made through transfer bus terminals available at the rear of the computer and at the problem board. All circuit connections were then made by wiring the program board.

The computer's operational amplifiers were high gain, solid state, chopper-stabilized dc amplifiers having excellent drift characteristics and a large dynamic output range. Care was taken in the measurements to avoid amplifier overload, drift, and parasitic oscillations, as well as

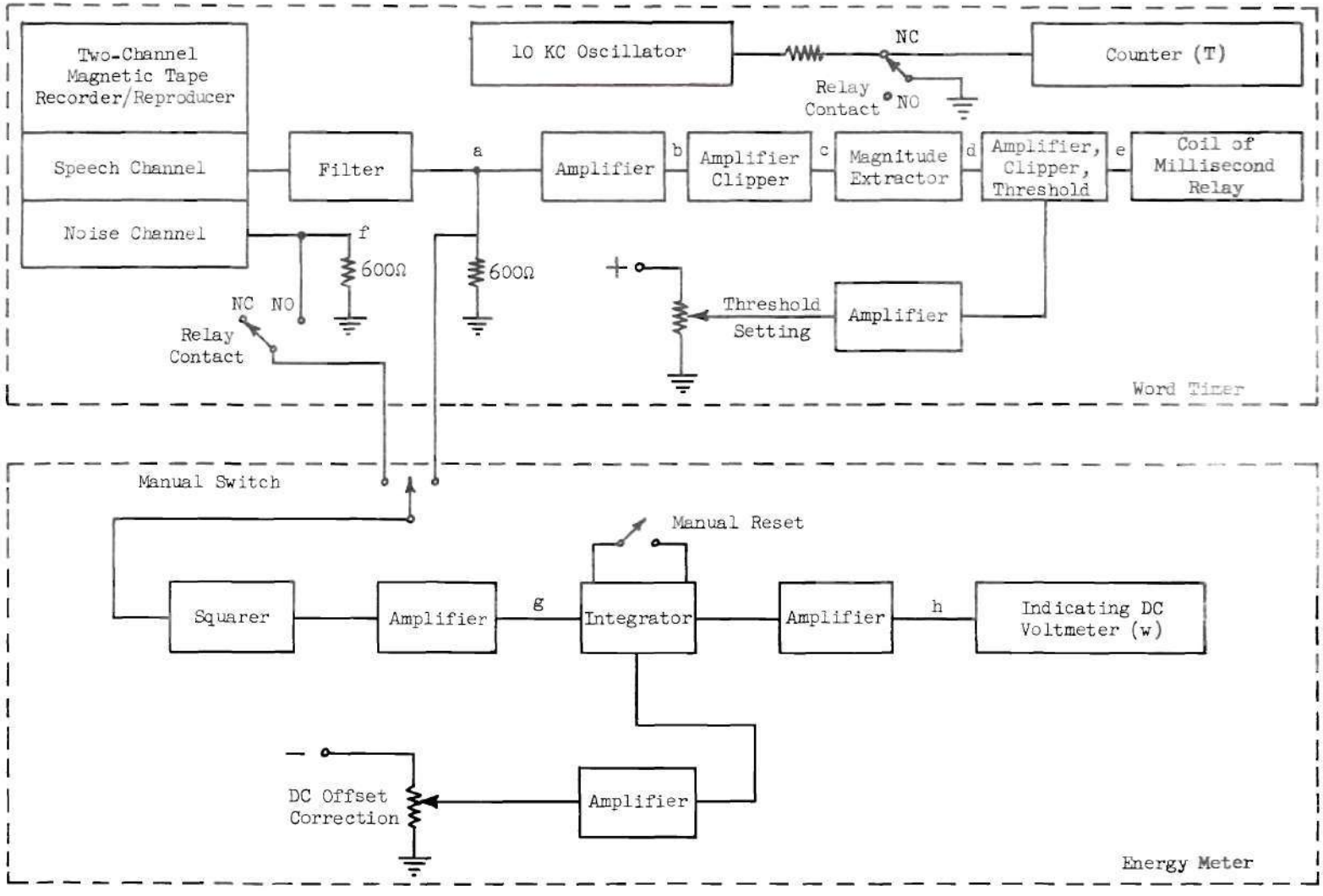


Figure 8. Block Diagram of Word Timer and Energy Meter.

ground loops in the circuitry. All external equipment was grounded to a single point in the computer ground system to aid in line frequency hum reduction. Waveforms at various points in the circuit were monitored with an oscilloscope having a long-persistence screen. A photograph of the equipment is shown in Figure 9.

Operation of the timer and energy meter will now be described, with reference to the voltage waveforms of Figure 10 and to the corresponding points in Figure 8. At the standard measurement points "a" and "f," highly reproducible speech and noise signals were available from the two-channel recorder. All measurements were made on the master tapes described earlier, these tapes being played at a standard level which was set by means of calibration tones recorded on each track of the tape.

Figure 9. Equipment for Measuring Word Duration and Energy.

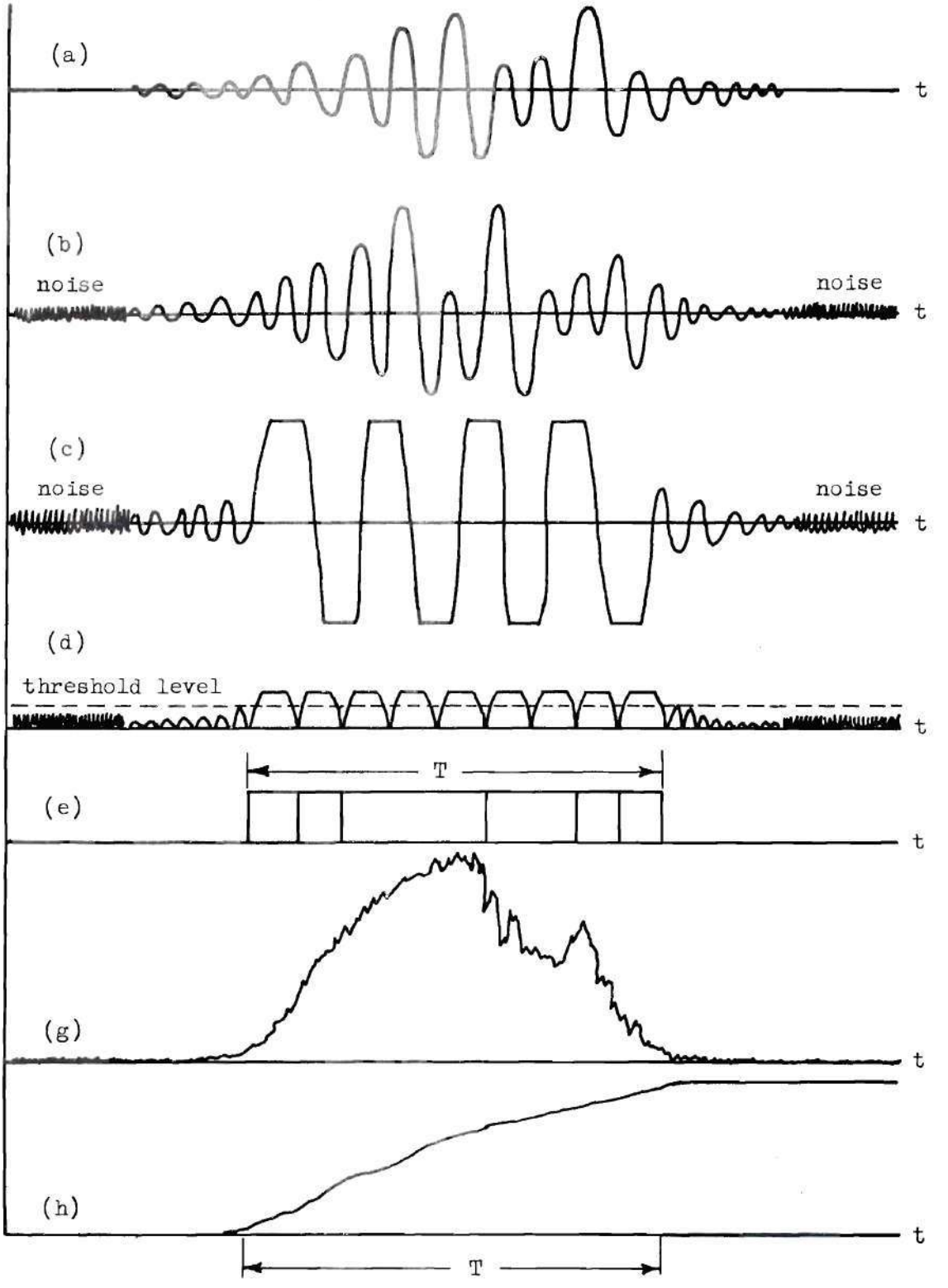


Figure 10. Voltage Waveforms at Various Points of Measurement System.

The peak speech to tape noise ratio was well in excess of 40 db, thus making available a signal essentially free of noise over the useful dynamic range of speech. Only a small amount of extraneous noises were inadvertently recorded on the tape between words, and these presented no difficulty in operation of the equipment. Both speech and noise signals were limited to eight kc in bandwidth, the former by means of the filter shown in Figure 8, the latter by an identical filter used while recording the noise.

The general principle of operation of the timer is similar to that of a previously reported device (6). The speech waveform of a single word at point "a" has the general form depicted in Figure 10. This, and other waveforms shown, are quite abbreviated as to number of zero crossings. They are also plotted as positive-going waves in all cases, using different vertical scales for convenience. The purpose of such simplified waveforms is to present, as clearly as possible, the basic features of equipment operation. After 20 db of amplification, the tape and amplifier noise are more pronounced, as shown for point "b." The word, after 20 db more amplification and approximately 30 db of symmetrical peak clipping, has the general form shown for point "c." In the magnitude extractor, two frequency-compensated operational amplifiers, interconnected in what is commonly known as a "V-circuit" configuration, provide full-wave rectification, resulting in the waveform shown for point "d."

The block following the V-circuit contains a high-gain amplifier/clipper stage having an adjustable base-clipping level, followed by a second amplifier/clipper. The first of these stages has a gain (below

clip level) of approximately 150 db, but its output is held at ground potential by the action of a fast diode clamp and by the application of an adjustable "threshold" voltage at the amplifier input. As a result, the amplifier provides an output only when the voltage at "d" exceeds the threshold value. When this occurs, the output increases rapidly to the peak clipping level set by a biased fast-diode clamp. For voltages in excess of the clip level of the first clipper this stage provides approximately 160 db of peak clipping, resulting, after further amplification and clipping in the last stage, in the waveform shown for point "e." At this point the noise, as well as word voltages on the order of the noise level, have been removed and the signal appears as a serrated rectangular pulse of length T (the "duration" of the word). Due to the large amount of peak clipping, the serrations, which are quite few in number for most words, have durations of considerably less than a millisecond, the actual value being set by amplifier bandwidth limitations rather than input rise time. The waveform at "e" is applied to a fast (millisecond) relay having two sets of SPDT contacts. Further electrical smoothing in the relay coil, plus inertial effects in the contact assembly, insured continuous chatterless contact closure during the interval T. For a few of the words this interval was composed of two or more distinct subintervals; these "holes" occurred at inter-phonemic transitions where the signal level dropped temporarily below the threshold value.

Closure of the relay contacts permitted a 10 kc clock signal from an oscillator to enter the digital counter during the time T, thus permitting the word duration to be read directly from the counter. A second set of contacts, actuated in synchronism with the timer, was available

for gating the energy meter input as described later.

The requirements for setting the timer sensitivity (i.e., the timer threshold), were that practically all of the word energy be contained in T, and that the threshold be far enough above noise level so that the timer was not unduly sensitive to spurious tape noises (such as breath noises made by the speaker during recording). The actual threshold selected was 46.5 millivolts, referred to the standard measurement point "a." This value was 40 db below the largest peak word voltage observed at this point, and resulted in only a very few spurious timer responses. These caused no difficulty in actual operation since they occurred between words and could easily be avoided by manually resetting the counter.

By operating the manual switch shown in Figure 8, the energy meter could be connected to either the speech or noise signal at standard measurement point "a" or "f," respectively. When measuring noise energy, it was necessary to gate the input to the squarer by means of the relay contact operated by the word timer; in this way only the T-second section of noise which was coincident with a word was measured. Except for the input gating the process of measuring this "noise word" energy was identical to that for words, and hence only the latter will be discussed. Squarer-integrator devices for measuring energy have previously been described (6,35,36), but in most of these the integrator was an RC type and in all but one case the integration time was fixed, though adjustable.

Due to the low background noise on the word channel of the tape, the energy meter had no detectable response during the 10-second interval between words, except for the small number of spurious noises recorded on the tape. As was true for the timer, the effect of such noises was easily

avoided, in this case by manually resetting the integrator just prior to a word. The speech waveform, after squaring and dc amplification, appeared as shown in Figure 10 for point "g"; this function of time is proportional to the instantaneous square of the word voltage. The squarer comprised an attenuator, amplifier, and a temperature-stabilized diode function generator, these being part of a Ballantine model 320 "true rms" voltmeter. The RC integrator and indicating meter of this instrument were disconnected during energy measurements. The function generator consisted of eight low-leakage germanium diodes interconnected with precision resistors and a regulated voltage source so as to obtain a nine-segment approximation to the ideal (parabolic) transfer function. The squaring error was less than one per cent over a 34 db range of input voltage and the attenuator setting was chosen such that this range coincided with the upper part of the 40 db speech range. The frequency range of this instrument is given as 5 cps to 500 kcps by the manufacturer; tests confirmed accurate squaring to at least 20 kc output frequency. The squarer was found to have a small residual dc output (with no input), but this was balanced out by adding an equivalent voltage of opposite polarity to the integrator input ("DC Offset Correction" in Figure 8).

The squared and amplified voltage is applied to an operational amplifier connected as an integrator and provided with a manual reset. The effective time constant of the integrator was 2.5×10^6 seconds, with the actual integration time being effectively the word duration T. The integrator output, after passing through a unity-gain buffer amplifier, appeared as shown for point "h." This voltage reached its maximum value at the end of T seconds and thereafter remained constant until the

integrator was reset, thus permitting an operator to read its value on an indicating dc voltmeter. The voltmeter reading was recorded as w.

Measured values of T for the 40-word master list ranged from 454.5 milliseconds for "ache" to 671.5 milliseconds for "rouse," and had an average value over the set of 546.7 milliseconds. Measured values of w ranged from 23.8 for "fig" to 67.3 for "gob," with an average value, over the set, of 35.64, while values of w_n for noise ranged from 35.7 to 52.5, with an average value of 42.1. The complete set of values, which are actually averages of four separate measurements, is shown in Table 2. From the four sets of measurements made, values of word and noise power were calculated using equations (1) and (4), and these averaged to obtain the values of p^i and p_n^i shown in Table 2. In addition, the mean word power $\bar{p}$ and mean noise power $\bar{p}_n$ were calculated from equations (2) and (5), as well as relative word and noise power in db with respect to these mean values, using equations (3) and (6). Finally, the individual SN^i ratios in db were calculated from equations (7) and (8), and the group ratio, SN, was calculated from equation (9). Only individual quantities are given in Table 2, but the group quantities $\bar{p}$, $\bar{p}_n$, and SN were found to be 64.9, 77.0, and -0.75 db, respectively. Note that these group quantities, as well as the set of SN^i, P^i, and P_n^i, are peculiar to this particular word-set, and are different for various subsets considered later. All quantities except T, P^i, and P_n^i are variable during an articulation test, values in the table being simply their reference values at the standard measurement point.

Fixed sources of error in measuring T include the difference in closure and release time of the relay (0.8 millisecond), round-off error

Table 2. Experimentally Determined Values of Duration, Energy, Power, and Signal/Noise Ratio for Words and Masking Noise in the Master Set

word	T^i	w^i	w_n^i	p^i	p_n^i	P^i	P_n^i	SN^i
ache	454.5	24.7	36.4	54.4	80.0	-0.77	+0.17	-1.68
bald	635.3	35.3	48.0	55.6	75.5	-0.67	-0.08	-1.33
bead	554.5	32.1	45.4	57.9	81.8	-0.50	+0.26	-1.50
cast	550.8	37.0	44.6	67.1	80.9	+0.14	+0.17	-0.81
check	489.5	24.9	35.7	50.8	72.9	-1.07	-0.25	-1.57
class	537.0	39.3	42.1	73.2	78.4	+0.52	+0.09	-0.30
crave	596.5	37.6	45.6	63.0	76.5	-0.13	-0.04	-0.94
crime	595.5	36.8	46.0	61.8	77.2	-0.22	0.00	-0.97
deck	499.3	29.4	39.7	58.8	79.5	-0.43	+0.13	-1.32
dill	625.0	31.5	46.8	50.4	74.8	-1.10	-0.13	-1.72
fame	519.0	36.5	40.5	70.4	77.9	+0.35	+0.04	-0.44
fig	469.5	23.8	37.2	50.8	79.2	-1.07	+0.13	-1.93
flush	518.0	50.8	37.3	98.1	71.9	+1.80	-0.32	+1.36
gnaw	614.8	37.8	47.5	61.5	77.3	-0.24	0.00	-0.99
gob	560.5	67.3	40.7	120.0	72.6	+2.67	-0.26	+2.19
hurl	605.8	37.5	47.3	61.9	78.0	-0.21	+0.04	-1.01
jam	585.5	33.7	44.2	57.5	75.5	-0.53	-0.08	-1.19
law	589.0	42.0	44.7	71.2	75.9	+0.40	-0.08	-0.28
leave	556.3	28.7	43.4	51.5	78.0	-1.01	+0.04	-1.81
lush	500.0	27.4	38.8	54.7	77.6	-0.74	+0.04	-1.52

Table 2. (Continued)

word	T^i	w^i	w_n^i	p^i	p_n^i	P^i	P_n^i	SN^i
muck	474.5	26.2	37.4	55.2	78.7	-0.71	+0.09	-1.54
neck	459.8	24.6	36.1	53.5	78.5	-0.84	+0.09	-1.66
nest	530.8	29.3	40.1	55.2	75.5	-0.71	-0.08	-1.36
path	484.0	33.3	37.6	68.8	77.6	+0.25	0.00	-0.53
please	552.5	36.2	40.7	65.4	73.6	+0.03	-0.21	-0.52
pulse	519.8	40.5	39.0	78.0	74.9	+0.80	-0.13	+0.17
rate	460.8	24.4	36.1	53.0	78.3	-0.88	+0.09	-1.70
rouse	671.5	64.6	52.5	96.2	78.1	+1.71	+0.09	+0.91
shout	610.0	30.8	46.9	50.5	76.9	-1.09	-0.01	-1.83
size	644.5	40.9	48.8	63.4	75.8	-0.10	-0.08	-0.78
stag	635.5	44.0	47.6	69.2	74.9	+0.28	-0.13	-0.34
take	489.3	26.0	38.2	53.1	78.0	-0.87	+0.04	-1.67
thrash	532.8	38.4	41.4	72.0	77.8	+0.45	+0.04	-0.34
toil	593.5	30.4	44.4	51.2	74.8	-1.04	-0.13	-1.64
turf	495.5	27.0	38.0	54.6	76.6	-0.75	-0.04	-1.48
vow	557.3	51.3	44.8	92.1	80.4	+1.52	+0.18	+0.59
wedge	581.5	34.8	43.7	59.9	75.2	-0.36	-0.10	-0.99
wharf	501.3	31.0	40.2	61.7	80.3	-0.22	+0.18	-1.14
who	507.0	26.5	39.4	52.3	77.7	-0.94	0.00	-1.72
why	511.3	51.3	38.6	100.4	75.5	+1.90	-0.08	+1.23

in the counter (a maximum of 0.6 millisecond), and errors in setting the clock frequency. Clock frequency was measured by an eput meter utilizing an internal crystal-controlled oscillator. In addition, this measurement was checked against a second eput meter; the two compared to within 0.1 per cent. Total maximum possible error due to these causes is only 0.53 per cent of the smallest measured value of T. Care was taken to minimize, as far as possible, random errors due to fluctuations in temperature, line voltage, and settings of playback level. Recorder reproduce heads were kept free of tape oxide particles by periodic cleaning, and intermittent oscillation of unused operational amplifiers was minimized by connecting output to input. Finally, in order to obtain a rough estimate of the repeatability of measurements, the set of four measured values of T and their mean value were examined for eight of the test words. Assuming a t-statistic, the 95 per cent confidence interval (37) about the mean, as a percentage of the mean, was calculated. The worst case found was ±1.5 per cent about the mean, implying that, for this case, the "true" mean is within ±1.5 per cent of the sample mean, with 0.95 probability.

The principal source of error in measuring w was the modified Ballantine voltmeter. The manufacturer specifies 3 per cent accuracy when the instrument is used as an indicating voltmeter for waveforms whose rms value, expressed as a fraction of full scale, is no more than five times the reciprocal of the crest factor. In order to apply this criterion, the rms value of each word was calculated from the power measurements, and the peak word voltages were measured by finding, for each word, the amount of attenuation necessary to reduce the peak just below the timer threshold. Crest factors were then computed. The results showed that no rms

value exceeded the full-scale value by more than seven per cent, and that the product of crest factor and fraction-of-full-scale rms value was below the manufacturer's upper limit of 5 except for one word. For this word ("gob"), the product was 5.8.

The squaring accuracy of the diode function generator was checked by measuring the dc component of the squared and amplified input voltage for various rms values of input. Ideally, these should be related by

$$E_{dc} = C E_{rms}^2 . \qquad (47)$$

The constant C was evaluated for both sinusoidal and band-limited gaussian noise inputs, by satisfying equation (47) at approximately half the maximum rms input used. The value obtained was the same in both cases, and was used to plot equation (47) as shown in Figure 11. Indicated also in this figure are the measured values of E_{dc} for various values of sinusoidal and noise inputs.

This graph shows only a 4 per cent full scale error for the squared sine wave, and indicates that when rms input does not exceed about 1 volt, excellent squaring exists over the range of input voltages. The largest rms voltage input for any of the test words was 1.07 volts. As a final check on squaring accuracy, the energy of each of 50 monosyllables was measured and the mean speech energy for the set was calculated. This was then repeated, with the speech input attenuated by 2 db. The resulting change in average measured speech energy was 2.09 db.

Measured values of energy showed somewhat less repeatability than values of T, especially for word energy. Considering 12 words selected on the basis of maximum spread in w and w_n, the confidence interval for

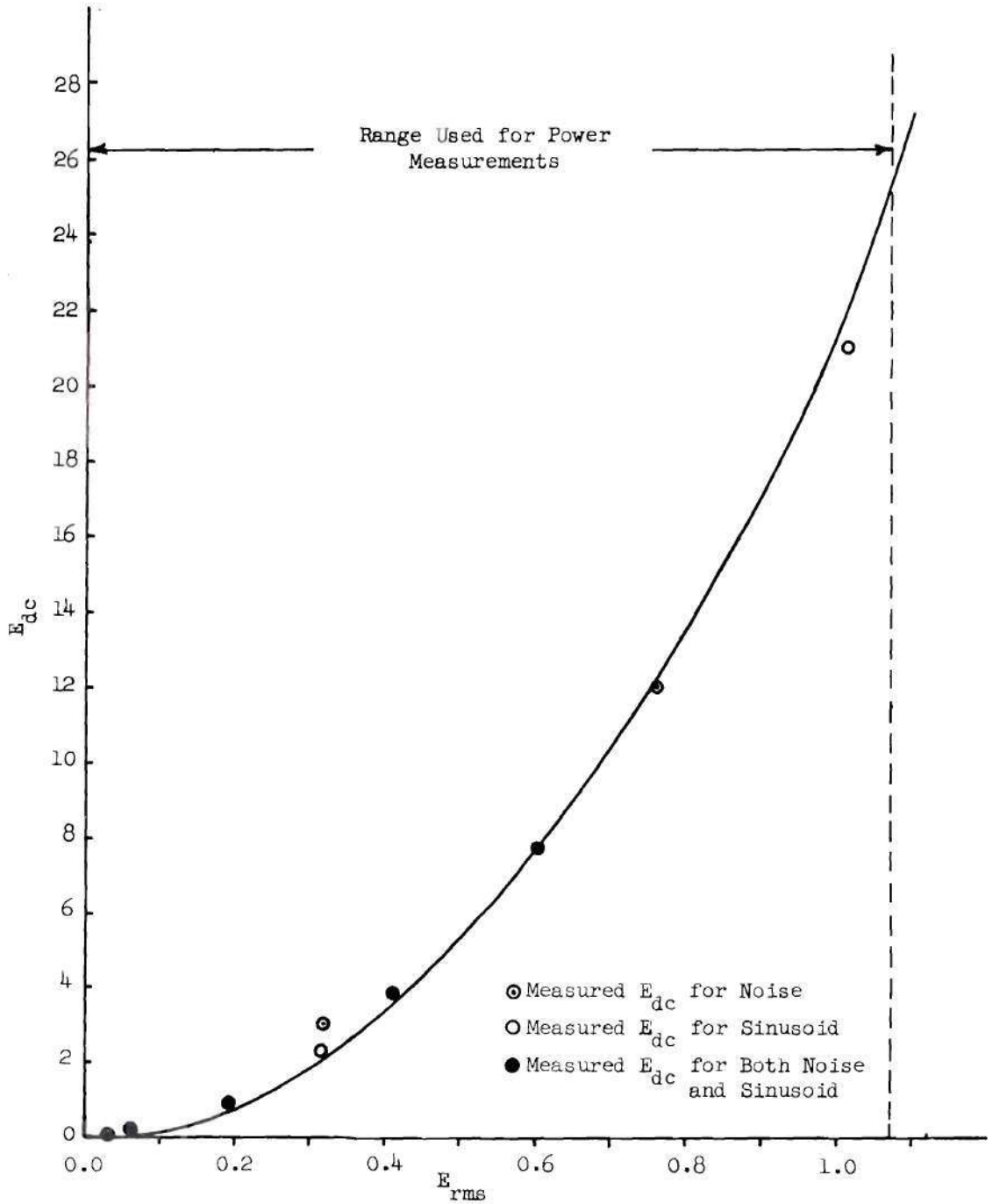


Figure 11. Illustration of Squaring Accuracy of Squarer.

the mean of four measured values of w, expressed in db with respect to the sample mean, was 0.86 db in the worst case for words, and 0.2 db in the worst case for noise. These calculations indicated that, with .95 probability, the true mean differed by no more than ±0.46 db from the sample mean in the case of word energy, and by no more than ±0.1 db in the case of noise energy. The repeatability of w_n measurements is clearly very good, being of the same order of magnitude as that for T, while the w measurements have a smaller, although acceptable, repeatability.

From the foregoing, it is apparent that the precision of values of both p^i and sn^i depends mainly on the repeatability of word energy measurements. Further consideration of data indicated that the values used for these quantities are probably correct to within 0.5 db.

As indicated in Chapter II, definitions of group SN ratio other than that given in equation (9) are possible. If one considers all the words of a set to be joined end-to-end, so as to form a continuous speech signal of duration $\sum_{i=1}^{N} T^i$, and if one also imagines all corresponding noise segments to be joined to form a noise signal of the same duration, then one might logically define the SN ratio of the set by

$$SN^{(1)} = 10 \log \left[\frac{\sum_{i=1}^{N} w^i}{\sum_{i=1}^{N} T^i} \div \frac{\sum_{i=1}^{N} w_n^i}{\sum_{i=1}^{N} T^i} \right], \qquad (48)$$

i.e., as the value in db of the ratio of word power in the continuous speech to noise power in the continuous noise. The expression (48) is

seen to be equivalent to

$$SN^{(1)} = 10 \log \frac{\overline{w}}{\overline{w}_n} , \tag{49}$$

where $\overline{w}$ and $\overline{w}_n$ are the mean word and noise energy for the set. Alternatively, one might define the group SN ratio as

$$SN^{(2)} = 10 \log \left[\frac{1}{N} \sum_{i=1}^{N} sn^i \right] , \tag{50}$$

or as

$$SN^{(3)} = \frac{1}{N} \sum_{i=1}^{N} SN^i , \tag{51}$$

where SN^i is defined by equation (8). Values of group SN ratio were calculated from equations (49), (50), and (51) for both the 40-word master set and for four 20-word subsets used in the tests. These values, along with those obtained by use of equation (9) and denoted by SN, are given in Table 3.

Table 3. Values of Group SN Computed from Various Definitions

Word Set	SN	$SN^{(1)}$	$SN^{(2)}$	$SN^{(3)}$
Master Set	-0.75	-0.73	-0.73	-0.85
Subset A	-0.76	-0.74	-0.75	-0.85
Subset B	-0.73	-0.72	-0.71	-0.85
Subset G	-0.71	-0.67	-0.70	-0.81
Subset H	-0.79	-0.78	-0.77	-0.89

It is clear from Table 3 that the various definitions give very nearly the same results, indicating that the choice of definition of group SN is immaterial, insofar as numerical values are concerned, for sets of at least 20 words. It is important, in measuring masking SN ratio, that the bandwidths of speech and noise be specified. In the work described here, both speech and noise were limited to the same (8 kc) bandwidth.

Articulation Tests

The general procedures used in the articulation tests are briefly described below, as are the test runs made on the 40-word master set. Tests made on 20-word subsets generally followed the same procedures; these subset tests are described in Chapter IV. A complete discussion of tape preparation is given in Appendix A, and a detailed description of test procedures, including listener training and effects of various subjective factors, is given in Appendix B. These appendices should be referred to for any details of the tests which are not discussed below.

The tests on the master word-set had two purposes, namely,

a. to obtain a conventional articulation curve, and

b. to obtain data for plotting the NMGF's.

No special tests were made to determine α or Δ for each word; these were determined, as described in the next section, from NMGF plots. In effect, the 40-word tests were made to do double-duty; by processing the test results in two different ways, both purpose "a" and purpose "b" were accomplished at the same time. In addition to convenience and reduced testing time, this procedure had the advantage of avoiding the residual learning effects to be expected from separate tests.

In cases where one is interested only in obtaining NMGF's, alternative procedures which save some testing time are possible. One way of doing this is to remove words from the list as soon as their NMGF's have been defined; the danger of thus decreasing the vocabulary size is that listeners who detect such changes will make choices from a smaller word-set and hence achieve higher scores. This effect of changing vocabulary size is well known (16, p. 77), and points up the fact that word parameters measured by subjective tests are not absolute but depend upon the size of the test vocabulary in general. Another way of obtaining NMGF's is to transmit one word to the listeners, at successively higher values of SN, until the word is heard on every transmission by all listeners, and to then repeat this process for other words in the set until all NMGFs have been defined. Each word must, of course, be transmitted in this way many times in order to obtain statistically valid mean scores at each of the SN ratios. A convenient way of doing this is to record each word on a separate loop of magnetic tape which is then played back continuously, with SN adjustments being made between occurrences of the word. The loops are then randomly selected and used in this way until each loop has been used a specified number of times. Each use of a given loop provides an NMGF for that word; the final curve is obtained by averaging these individual curves at each point. Some preliminary tests made in this way indicated a serious fault, namely, that the listener tends to be biased in subsequent responses on the same loop once he has understood, or thinks he has understood, the word on that loop.

The actual procedure used was to play an entire tape, containing 40 words, at some initial SN ratio, and to repeat this process, using

different tapes having the same words but in different orders, at steadily increasing values of SN until 10 such "runs" had been made. Each run provides a single point on the articulation curve, while each collection of 10 runs, termed a "repetition," describes a complete 10-point curve. This process was repeated until 10 repetitions had been made, following which the score-sheet data was processed to yield the desired curves. The tapes used in these listener tests were transcribed from the master tapes which had been used for power measurements, i.e., they were second-generation copies. By using great care in the transcription process, eight copies were obtained which were practically identical to the master. Minor differences in these tapes were averaged out by randomizing the order of tapes from run to run such that each tape was played at least once at every SN ratio. The initial and final runs of each repetition were made at SN ratios outside the range of interest; specifically, they were played at SN values of-8.75 db and -26.75 db. Tapes used for these points were somewhat less perfect copies of the master tape than the others, and these runs are not included in the final data. The eight "good" copies were used to cover the range -10.75 db to -24.75 db in two-db steps; these values of SN corresponded to integral settings of a step attenuator when the tapes were played back at standard level.

In addition to preliminary "practice words," the spaced test words of Table 1, and (on a separate track) continuous noise, each tape contained auditory "cues" superimposed on the noise. These were in the form of a short burst of two kc tone occurring approximately three seconds before each word, and served to alert the listeners and thus prevent words being missed due to inattention or momentary distraction.

There were several criteria used for selecting the 40 test words from the PB lists (all 1000 PB words were originally recorded on the master tapes). First, the four best masters, each containing one PB list, were selected, using the criteria of good pronounciation and voice quality. Tapes having slight defects such as momentary "drop-outs" due to irregularities in the oxide coating were discarded during this process. Next, the tapes were monitored aurally and visually, using headphones and a VU meter, for uniform, accurately timed cues and for constancy of noise level. Finally, 39 words from PB 3 and 1 word from PB 20 were chosen. The PB 20 word was chosen to increase the spread in word power for the set; it had one of the largest measured values of power. The remaining words were selected with a view toward investigating the effect of common-vowel groupings in a word set. The final set contained groups of 6, 5, 4, 3, 2, and 1 words, all words within each group having the same vowel sound. There were, respectively, 1, 2, 2, 3, 2, and 3 groups having the number of common-vowel words specified above. Thus, for example, there were two 5-word groups, three 3-word groups, three 1-word groups, and one 6-word group.

The test equipment was set up in an acoustically treated room, in which were individual tables for the listeners, each equipped with a pair of high-quality headphones and an adjustable level control. Provision was made for an operator who distributed and collected score sheets and monitored all tests. In addition, the operator calibrated the equipment and set the recorder play-back levels to standard values. A block diagram of the equipment is shown in Figure 12.

The recorder was the same one used in making the original masters

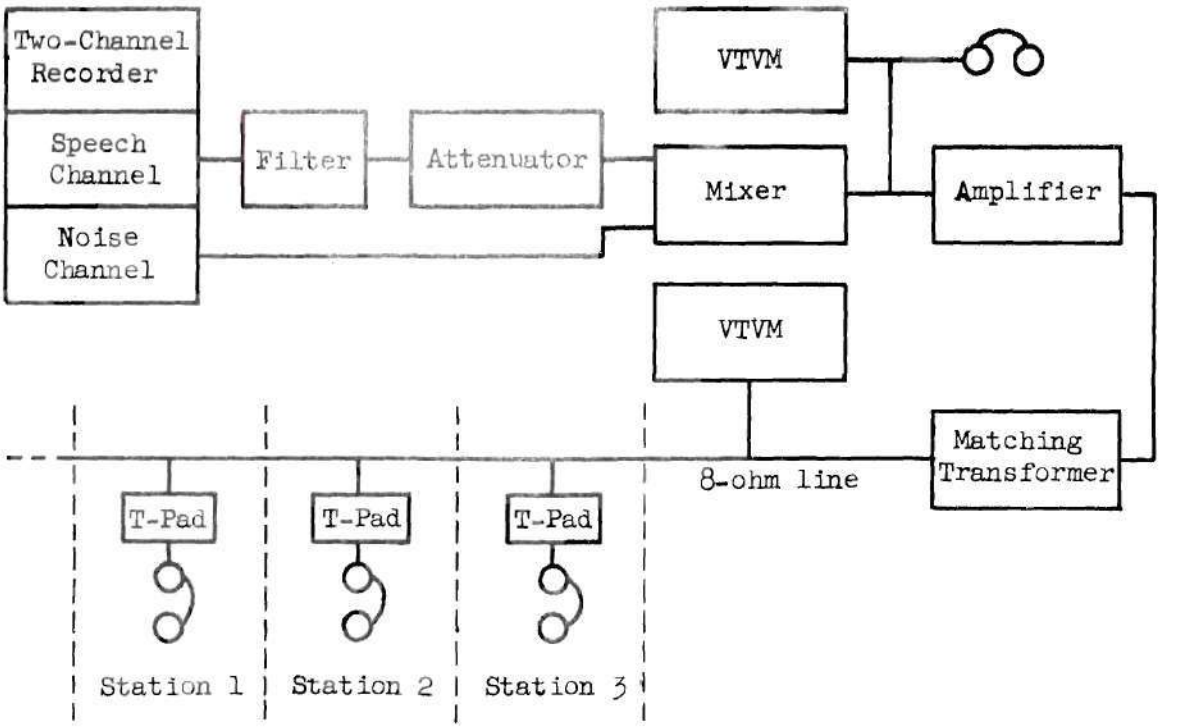


Figure 12. Block Diagram of Articulation Test Equipment.

and in the power measurements. It was also one of two identical recorders used in the transcription process for making copies. The step-attenuator provided a means of adjusting the SN ratio by attenuation of the speech. The filter and attenuator both had 600 ohms iterative impedance, the proper termination for the attenuator, as well as for the noise channel output, being provided by a resistive bridge-type mixer. The speech and noise were linearly added in this mixer, which had an insertion loss of 6.28 db to both speech and noise and an isolation between inputs of 52 db. The mixer output was monitored

with a vacuum tube voltmeter and high-impedance crystal phones, the VTVM being used to set the recorder's playback level. The noise-masked speech was amplified in a high-quality audio amplifier possessing flat frequency response and low distortion over the signal bandwidth, and whose output was matched to an eight-ohm line by means of a high-quality matching transformer. Only 3 of 6 available listening stations were used; each consisted of a set of 50-ohm dynamic binaural phones connected to the line by means of an adjustable 50-ohm T-pad. Dummy 50-ohm resistive loads were used to terminate the pads on the unused stations. The amplifier gain was calibrated by means of a second VTVM connected to the line.

Three undergraduate students (JB, WN, and NS) possessing normal hearing acuity and having previous listening experience were trained for use as listeners, using the master list (see Appendix B for details of the training program and the instructions given the listeners). Listener response was not strictly forced-choice, the listeners being allowed to indicate on the score sheet those words which could not be heard or which were too unintelligible to permit a reasonable guess as to their identity.

The completed score sheets were graded by the operator, who compared the response to a randomization key available only to him. Each response was marked as correct or incorrect, and word scores in per cent recorded for each run.

Following the above procedures, 10 repetitions were made, involving 4,000 word transmissions to each of the three listeners. The resulting data was tabulated in a form permitting the rapid determination of NMGF points for individual words and listeners. The scores of each listener were also averaged over the 10 runs at each SN ratio in order to obtain

eight-point, master-set articulation curves. These curves were then averaged over the listeners to obtain a "team curve" for the master set. Some NMGF and articulation curves are presented in following sections.

All data from the articulation tests were originally in terms of attenuator settings. Since these settings represent the number of db by which the SN ratios are decreased from their values at the standard recorder output level, the values of SN^i and SN corresponding to a given attenuator setting of $|X|$db are easily calculated from the measured values at standard level. Thus the signal/noise ratio at the phones is given, for the i^{th} word, by

$$SN^i = \left[SN^i \right]_{measured} + X \ , \tag{52}$$

and the group signal/noise ratio is

$$SN = \left[SN \right]_{measured} + X \ , \tag{53}$$

where X is the ratio, in db, of attenuator output to input (a negative number in all cases).

Noise-Masked Gain Functions and Their Parameters

Each point on the NMGF for a particular word and listener was calculated as follows:

1. The data was examined for that word and that listener at the lowest of the eight test values of SN ratio, and the number of correct responses, out of the 10 repetitions, was determined.

2. The fraction of correct responses was determined by dividing the number of correct responses by 10, resulting in the fractional word

score at the SN value in question.

3. The above steps were repeated for each of the remaining seven values of SN ratio, thus determining eight points on the NMGF which were then plotted.

4. Steps 1 through 3 were repeated for each of the words and for each of the listeners, resulting in 120 collections of NMGF points.

It is clear from the foregoing that the calculated points are simply an alternative way of presenting the same data which goes into the conventional articulation curve.

At this stage, the abscissa for each point was still given in terms of X, the "gain" from input to output of the attenuator. The independent variable used for plotting depends upon the use to be made of the NMGF. If it is to be compared with other gain functions as to location along a signal/noise ratio axis, then the individual SN^i of that word is the logical choice. If the prediction schemes of Chapter II are to be applied to a set of these functions, then SN, for that set, is the logical choice, bearing in mind that SN depends upon the particular words in the set. Since the functions were to be used for both purposes, including application of prediction schemes to five different word-sets, the data was left in terms of X until further processing could be done. Note that the shape of the functions and their spread depend in no way upon the choice of independent variable, furthermore, values of SN^i and SN are readily obtained from values of X, by using equations (52) and (53).

Several typical and atypical noise-masked gain functions are shown in Figure 13, with SN^i as the independent variable. As can be seen from the figure, when smooth curves can be fitted visually to the sets of

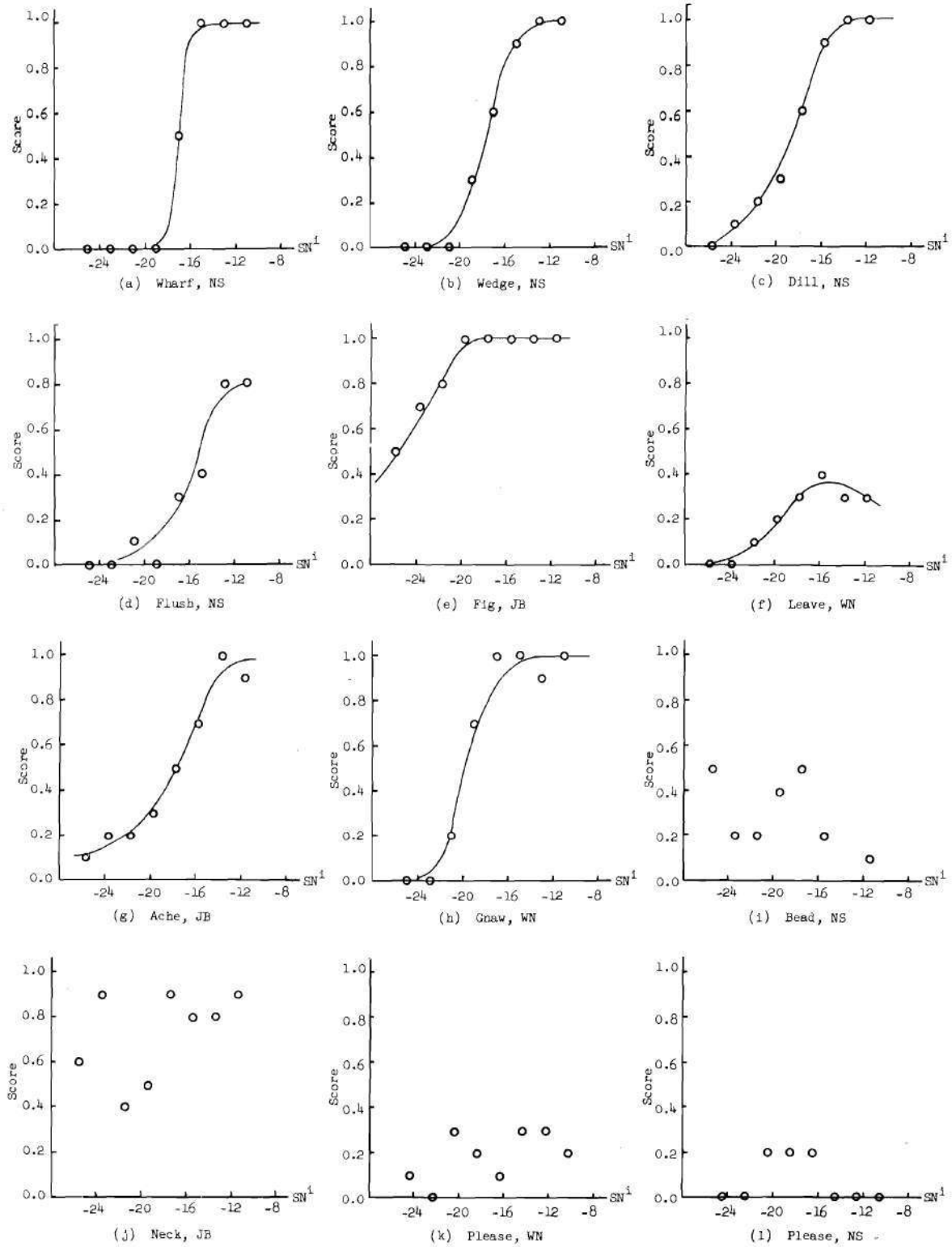


Figure 13. Examples of Noise-Masked Gain Functions.

points, they are quite varied in shape. Curves (a) through (e), plus (g) and (h), are more typical of the results than the others. Curves of this general type occurred in about 85 per cent of the cases, and are characterized by fairly smooth and monotone non-decreasing variation from low to high scores, with the exception of at most one point. The curves of (a), (b), (c), and (d) illustrate the case where essentially the entire function is defined by the eight attenuation values used to cover the test range, whereas the curves of (e) and (f) illustrate functions which are only partially defined by the test points. Curves in (a), (b), and (c) show, in that order, increasing values of SN spread between 0 and 1.0 fractional score, while curves (d) and (e) exhibit widely differing values of SN^i at their 0.5 levels. Various atypical curves exhibited anomalies such as definite reversals in direction, negative or very small slopes over the entire test region, and lack of any definite trend. The latter two types of anomaly are illustrated by the data points of (i), (j), (k) and (l).

It is clear that a three-segment piece-wise linear approximation, as required by one of the prediction schemes of Chapter II, cannot provide a good fit to a collection of points such as shown in part (j). On the other hand, such extreme cases were relatively rare and thus, in order to test the prediction scheme, all sets of points were fitted by such linear approximations. This was done by using an arbitrary rule, applied uniformly to all sets of points, to divide the points into three contiguous regions: a region in which the curve could be approximated by a horizontal line segment through the origin, a region in which the curve could be approximated by a straight line segment of non-zero slope, and a

region in which the curve could be approximated by a horizontal line segment through the 1.0 value of score. A complete set of such regions did not, of course, exist for all words. The linear approximation was then obtained as follows:

1. The points to be approximated by a line of non-zero slope were tabulated for each word and each listener, with values of attenuator gain X as the independent variable.

2. A straight line of slope m, X-intercept X_o, and abscissa X_{50} at the 0.5 value of score, was fitted to the points by the method of least squares, employing a high-speed digital computer for the calculations. The computer also calculated Δ, the abscissa interval in which the regression line spanned unit distance along the ordinate scale.

3. Using the results of step 2, the least-squares line was drawn through the scatter of points and terminated at score values of 0.0 and 1.0.

4. Horizontal straight-line segments were drawn to the end-points of the regression line segment.

In the cases where part of the curve lay outside the test region, steps 3 and 4 were not completed, i.e., only that part of the linear approximation lying in the test region was drawn. Some of the resulting curves are shown in Figure 14, with dashed-line extensions of the curves drawn beyond the test region in some cases.

Approximations for which the sloping line-segment lay essentially within the test region, shown in part (a) of Figure 14, were the most numerous, comprising about 68 per cent of the curves. To the extent that a good fit is obtained, values of X_{50} and Δ read from such lines give good

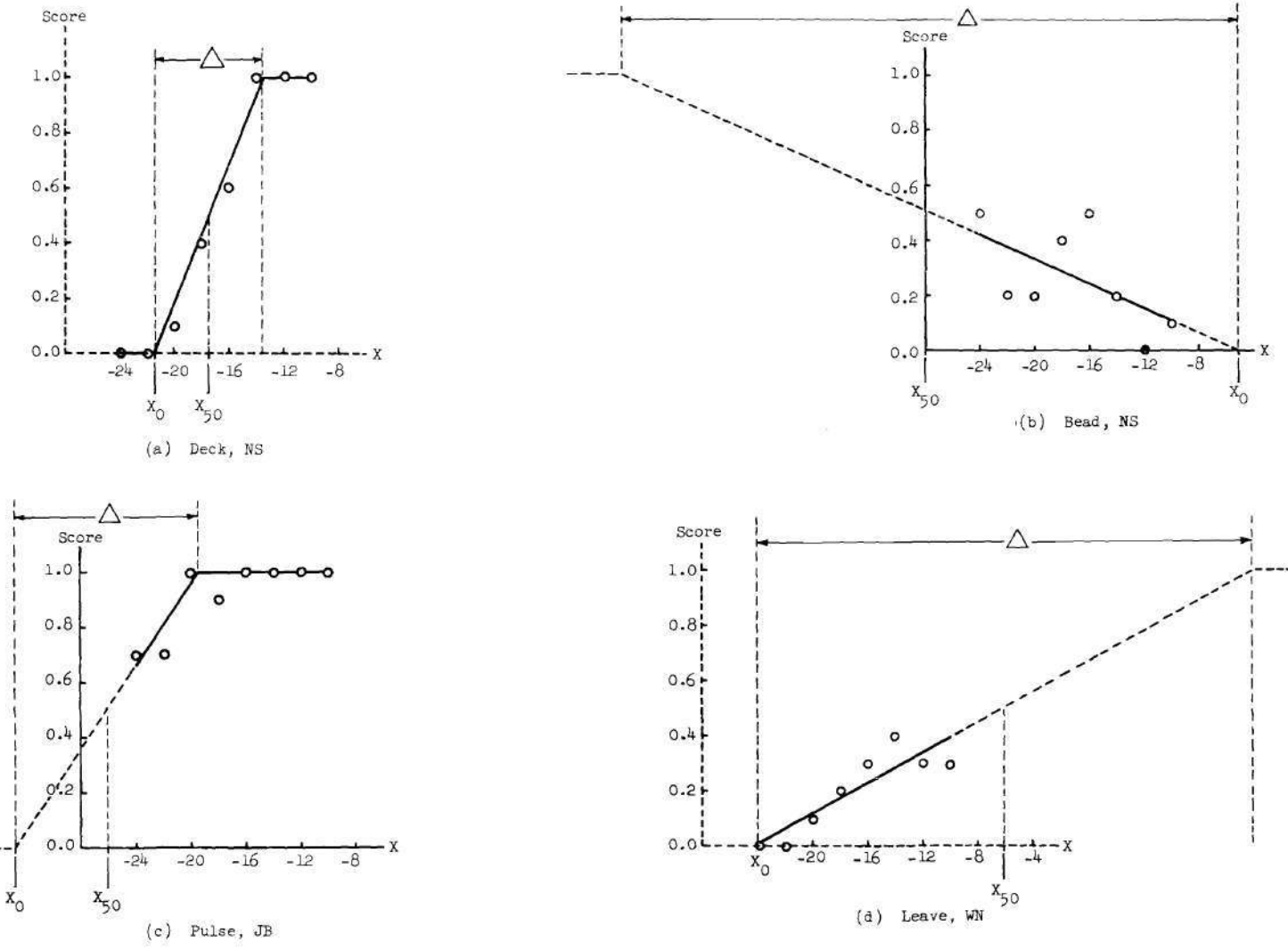


Figure 14. Linear Approximations to Noise-Masked Gain Functions.

measures of the threshold and spread of the actual curve. After computing α from X_{50}, these "valid" values of α and Δ can be used to compare words (for a given listener) or listeners (for a given word), as discussed in Chapter V. Of the remaining linear approximations, all but a few were sufficiently well defined in the test region to yield valid values of α and Δ. Examples of linear approximations yielding non-valid values of threshold and spread are shown in parts (b), (c), and (d) of Figure 14. Such "non-valid" cases comprised only about 9 per cent of all curves.

For linear approximations having negative slopes, such as in part (b) of Figure 14, the value of α computed from X_{50} is clearly not a "threshold" in the usual sense, nor is Δ a valid measure of the spread. An ambiguity as to the nature of the curve outside the test range, evident in parts (c) and (d) of the figure, causes the values of α and Δ determined from the extended line to be non-valid estimates of threshold and spread for the corresponding word-listener combinations. In four cases, the values of α determined from the approximation was judged to be valid, while the corresponding Δ's were judged to be non-valid. In all, 113 valid values of α and 109 valid values of Δ occurred out of the 120 curves.

It is important to note that the words "valid" and "non-valid," as used above, apply only to the use of values of α and Δ as estimates of the actual threshold and spread. For other uses, such non-valid values may be perfectly acceptable. In particular, it is shown later that the use of such values in the prediction schemes yields good results. This is to be expected if prediction is confined to the test region, and if, in this region, the linear approximation is a good estimate of the NMGF.

Such values of α and Δ are then viewed not as estimates of threshold and spread, but merely as constants describing a straight line. A method of handling negative-slope lines in the prediction schemes has already been given in Chapter II.

For the prediction scheme requiring only values of α, or of the related variable β, the NMGF's are, in effect, replaced by a unit step at α (or at β or at X_{50}, depending on the variable being used), as shown in Figure 15. Parts (a) and (b) of this figure illustrate the two possible results, namely, a "conventional" threshold and a "reversed" threshold. Again, the reversed steps, as well as some of the others, do not represent valid estimates of threshold, but are acceptable, for the prediction scheme, as mathematical approximations to the NMGF's.

From values of X_{50} furnished by the computer, the values of α^i were calculated, using equation (52) and the fact that α^i is simply the

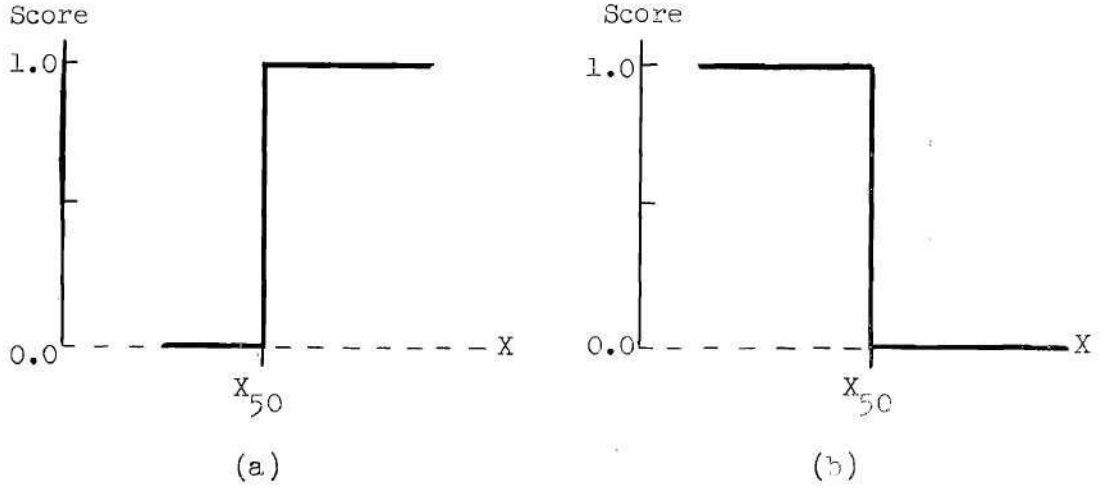


Figure 15. Step Approximations to NMGF's.

value of SN^i when X is equal to X_{50}. The β^i can be calculated from equation (17), but a simpler method is to make use of the fact that

$$\beta^i = \alpha^i - P^i + P_n^i = \alpha^i - SN^i + SN \qquad (54)$$

$$= (SN^i + X_{50}) - SN^i + SN = X_{50} + SN \ ,$$

where SN is the observed group signal noise ratio at the standard measurement point. Since SN differs for different word-sets, the calculations must be repeated for each distinct set.

Finally, the values of X_{50}, α^i, β^i and Δ^i were tabulated for each listener and for each of the words of the master set, using SN for that set. The results are given in Table 4.

Tables 2 and 4 contain all of the basic word parameters for the master set. It has been shown, in Chapter II, that the articulation curve is expressible mathematically in terms of these parameters. Before illustrating the application of this technique, two factors affecting the accuracy of the linear approximations will be briefly considered.

One source of error in the fitted lines is the error in determining the points defining the NMGF. This error is greatest near the 50 per cent level of intelligibility, i.e., the confidence which one has in an experimentally determined NMGF point is least when the ordinate of that point is 0.5. The ordinate observed at a given SN ratio can be viewed as an estimate of the probability θ of success (word understood) at that SN ratio. The confidence interval for θ is an interval such that one can state, with a given probability ρ, that θ lies within the interval. This interval was determined (37), for $\rho = 0.95$, in the case where the experi-

Table 4. Noise-Masked Gain Function Parameters

word	Listener JB				Listener WN				Listener NS			
	X_{50}	α^i	β^i	Δ^i	X_{50}	α^i	β^i	Δ^i	X_{50}	α^i	β^i	Δ^i
ache	-17.00	-18.68	-17.75	14.00	-14.53	-16.21	-15.28	4.00	-17.56	-19.24	-18.31	11.11
bald	-16.68	-18.01	-17.43	9.46	-14.00	-15.33	-14.75	8.00	-14.60	-15.93	-15.35	10.00
bead	-5.69	-7.19	-6.44	129.23	-71.60	-73.10	-72.35	168.00	-27.78	-29.28	-28.53	45.41
cast	-20.46	-21.27	-21.21	7.69	-18.82	-19.63	-19.57	10.77	-19.00	-19.81	-19.75	10.61
check	-15.73	-17.30	-16.48	4.00	-13.00	-14.57	-13.75	5.88	-14.13	-15.70	-14.88	4.00
class	-11.85	-12.15	-12.60	7.69	-15.80	-16.10	-16.55	12.07	-10.62	-10.92	-11.37	15.38
crave	-17.76	-18.60	-18.51	6.06	-19.73	-20.57	-20.48	4.00	-17.60	-18.44	-18.35	4.00
crime	-17.94	-18.90	-18.69	6.25	-20.00	-20.96	-20.75	8.33	-19.45	-20.41	-20.20	9.09
deck	-19.28	-20.60	-20.03	5.56	-17.87	-19.19	-18.62	10.45	-17.36	-18.68	-18.11	8.00
dill	-15.15	-16.87	-15.90	9.09	-21.07	-22.79	-21.82	13.33	-17.36	-19.08	-18.11	11.20
fame	-14.24	14.68	-14.99	9.09	-18.63	-19.07	-19.38	10.98	-14.57	-15.01	-15.32	5.71
fig	-24.13	-26.06	-24.88	12.50	-16.38	-18.31	-17.13	16.63	-13.33	-15.26	-14.08	8.33
flush	-19.12	-17.76	-19.87	8.00	-16.00	-14.64	-16.75	7.14	-13.62	-12.26	-14.37	9.52
gnaw	-22.89	-23.88	-23.64	6.67	-18.86	-19.85	-19.61	5.71	-15.59	-16.58	-16.34	14.36
gob	-17.15	-14.96	-17.90	6.06	-18.00	-15.81	-18.75	4.00	-15.94	-13.75	-16.69	6.25
hurl	-20.38	-21.39	-21.13	11.86	-18.80	-19.81	-19.55	8.00	-17.44	-18.45	-18.19	8.86
jam	-12.40	-13.59	-13.15	4.00	-11.56	-12.75	-12.31	4.44	-12.86	-14.05	-13.61	5.71
law	-15.52	-15.80	-16.27	8.00	-16.51	-16.79	-17.26	9.86	-17.74	-18.02	-18.49	10.77
leave	-16.40	-18.21	-17.15	4.00	-5.87	-7.68	-6.62	37.33	-16.00	-17.81	-16.75	7.41
lush	-15.90	-17.42	-16.65	9.52	-13.67	-15.19	-14.42	10.00	-12.86	-14.38	-13.61	5.71

Table 4. (Continued)

word	Listener JB				Listener WN				Listener NS			
	X_{50}	α^i	β^i	Δ^i	X_{50}	α^i	β^i	Δ^i	X_{50}	α^i	β^i	Δ^i
muck	-18.39	-19.93	-19.14	9.21	-19.15	-20.69	-19.90	6.06	-22.27	-23.81	-23.02	14.00
neck	-28.81	-30.47	-29.56	52.50	-21.09	-22.75	-21.84	9.09	-17.97	-19.63	-18.72	10.29
nest	-12.86	-14.22	-13.61	5.41	-11.24	-12.60	-11.99	5.71	-11.93	-13.29	-12.68	7.14
path	-14.40	-14.93	-15.15	4.00	-14.13	-14.66	-14.88	6.45	-14.17	-14.70	-14.92	8.70
please	-13.52	-14.04	-14.27	19.31	+8.00	+7.48	+7.25	80.00	-136.00	-136.52	-136.75	280.00
pulse	-26.00	-25.83	-26.75	13.33	-22.30	-22.13	-23.05	7.41	-19.15	-18.98	-19.90	6.06
rate	-15.76	-17.46	-16.51	6.06	-14.33	-16.03	-15.08	8.33	-13.38	-15.08	-14.13	7.69
rouse	-15.67	-14.76	-16.42	6.67	-15.87	-14.96	-16.62	6.45	-15.90	-14.99	-16.65	9.52
shout	-13.87	-15.70	-14.62	6.45	-15.00	-16.83	-15.75	5.56	-14.67	-16.50	-15.42	8.33
size	-16.61	-17.39	-17.36	8.70	-17.52	-18.30	-18.27	8.00	-15.29	-16.07	-16.04	5.88
stag	-15.16	-15.50	-15.91	6.45	-17.14	-17.48	-17.89	5.71	-17.00	-17.34	-17.75	6.25
take	-18.46	-20.13	-19.21	7.69	-21.19	-22.86	-21.94	7.41	-17.26	-18.93	-18.01	20.74
thrash	-13.80	-14.14	-14.55	10.00	-17.84	-18.18	-18.59	8.00	-17.17	-17.51	-17.92	13.44
toil	-19.52	-21.16	-20.27	10.45	-17.33	-18.97	-18.08	11.67	-17.04	-18.68	-17.79	8.00
turf	-16.13	-17.61	-16.88	4.00	-24.40	-25.88	-25.15	8.00	-16.44	-17.92	-17.19	7.41
vow	-13.73	-13.14	-14.48	4.00	-16.05	-15.46	-16.80	9.46	-51.00	-50.41	-51.75	100.00
wedge	-21.69	-22.68	-22.44	13.46	-18.42	-19.41	-19.17	14.24	-16.46	-17.45	-17.21	7.69
wharf	-18.67	-19.81	-19.42	9.52	-18.86	-20.00	-19.61	5.71	-16.00	-17.14	-16.75	4.00
who	-16.00	-17.72	-16.75	7.69	-24.33	-26.05	-25.08	10.00	-21.45	-23.17	-22.20	9.09
why	-21.33	-20.10	-22.08	6.67	-25.00	-23.77	-25.75	10.00	-21.00	-19.77	-21.75	6.67

mentally found ordinate was 0.5, i.e., where the word was understood in 5 out of the 10 repetitions made. This 95 per cent confidence interval extended from 0.18 to 0.82, thus permitting one to state, with 95 per cent confidence, that when the measured value is 0.5, the "true" value lies between 0.18 and 0.82. It is clear that a large amount of uncertainty exists in this worst case. Furthermore, the number of repetitions which must be made to decrease the confidence interval to a reasonably small value is prohibitively large. To obtain a 95 per cent confidence interval which extends from 0.45 to 0.55 (±10 per cent spread about the observed value) would require over 250 repetitions, almost an impossibility from the standpoint of maintaining stable test conditions for tests on 40 words at eight SN ratios. Fortunately, the confidence in a value of score predicted from 40 NMGF points is considerably higher than the confidence in one point.

A second source of error arises from the fact that a straight line is not a perfect fit to the set of NMGF points. A measure of the "linearity" of the NMGF, i.e., a measure of the extent to which score is linearly dependent on SN ratio, is given by the linear correlation coefficient r (38). These correlation coefficients, when computed for each of the 120 regression lines, were found to range in magnitude from 0.141 to 1.000. A histogram for $|r|$, shown in Figure 16, reveals that about 78 per cent of the correlation coefficients exceed 0.90 in magnitude, indicating generally good linear approximations.

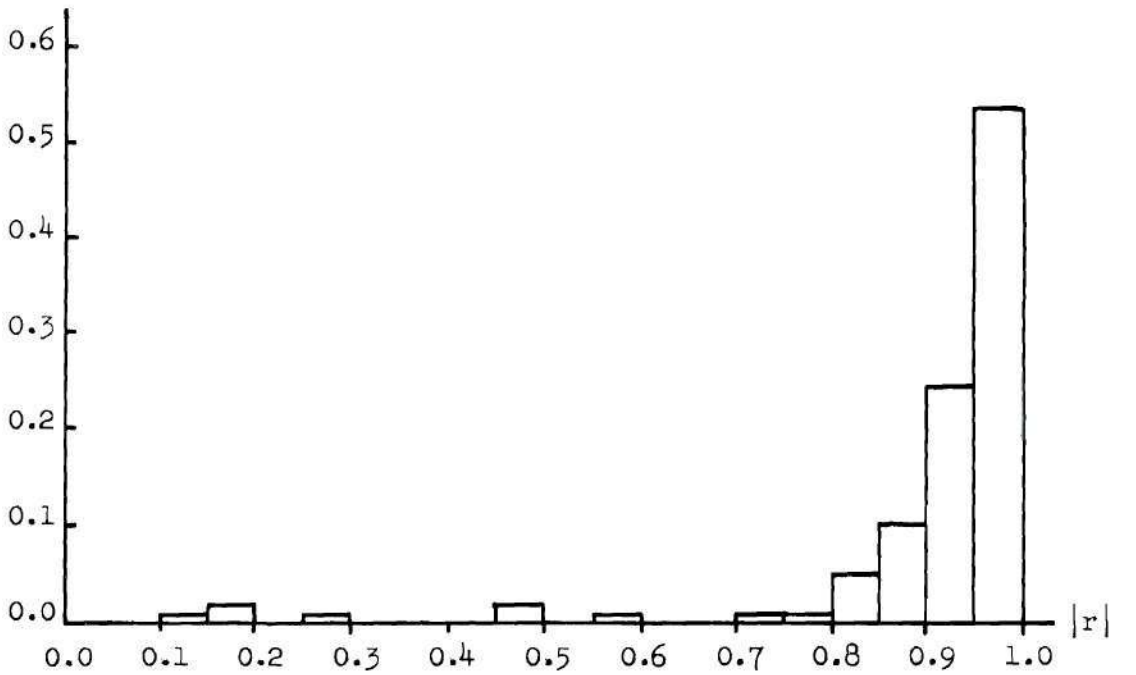


Figure 16. Histogram for Magnitude of Linear Correlation Coefficient.

CHAPTER IV

PREDICTION AND SHAPING OF ARTICULATION CURVES

Application of Prediction Schemes to Master Set

The two prediction schemes of Chapter II, together with the tabulated word parameters of Table 4, provide the necessary means for "predicting" the articulation curve for the 40-word master set. The resulting curves are not truly predictions, in the usual sense of the word, but merely represent alternate ways of using the basic articulation test data for obtaining a classical articulation curve. When these curves are compared with curves obtained in the conventional manner, one would expect close agreement if the following hypotheses are true:

(a) The prediction schemes are sound, and

(b) The word parameters α (or β) and Δ used in the prediction contain essentially all of the information to be found in the conventional curve.

In a sense, then, applying these schemes to the master set tests the validity of the word parameters as a means of representing the intelligibility of a set of words. The use of thresholds alone, and the use of both thresholds and spreads, represent two different degrees of approximation to the NMGF's. One would expect the former, in which an NMGF is replaced by a single-step function, to result in a poorer "prediction" than the latter, in which an NMGF is replaced by straight line segments.

The first step in applying the step-approximation scheme is to arrange the values of β^i from Table 4 in ascending order, at the same time

removing, and arranging in ascending order, any values corresponding to words having "reversed" step approximations (negative-slope NMGF's). When this was done for listener JB, application of equation (34) to the two sets of values yielded the typical monotone-increasing "stair-step" curve shown in Figure 17. In this and other following curves, the word score has been expressed in per cent, rather than in fractional values obtained from the prediction equations. The abscissa of the curve is group signal/noise ratio, SN, for the master set.

The linear approximation scheme was next applied, using equation (29) and points from the linear approximations to the NMGF's. This was done only at the eight test values of SN, not only because of the considerable time involved in calculating each point (these calculations were made by hand), but also because experimentally-obtained scores were available for comparison only at these points. The points from the linear prediction scheme, as well as those from the experimental articulation curve, are plotted for listener JB in Figure 17. Note that the experimental points could have been obtained from the collection of NMGF points calculated as described in Chapter III, since this collection of points is simply another way of expressing the articulation curve, and hence these points may be viewed as applying equation (29) to the NMGF's with no approximation.

From a comparison of the three values of score at each of the test points of Figure 17, it is clear that the errors resulting from the linear and step approximation of the NMGF's are small. The maximum error for the step approximation (with respect to the conventional score) occurs at the highest value of SN and is only about five percentage points

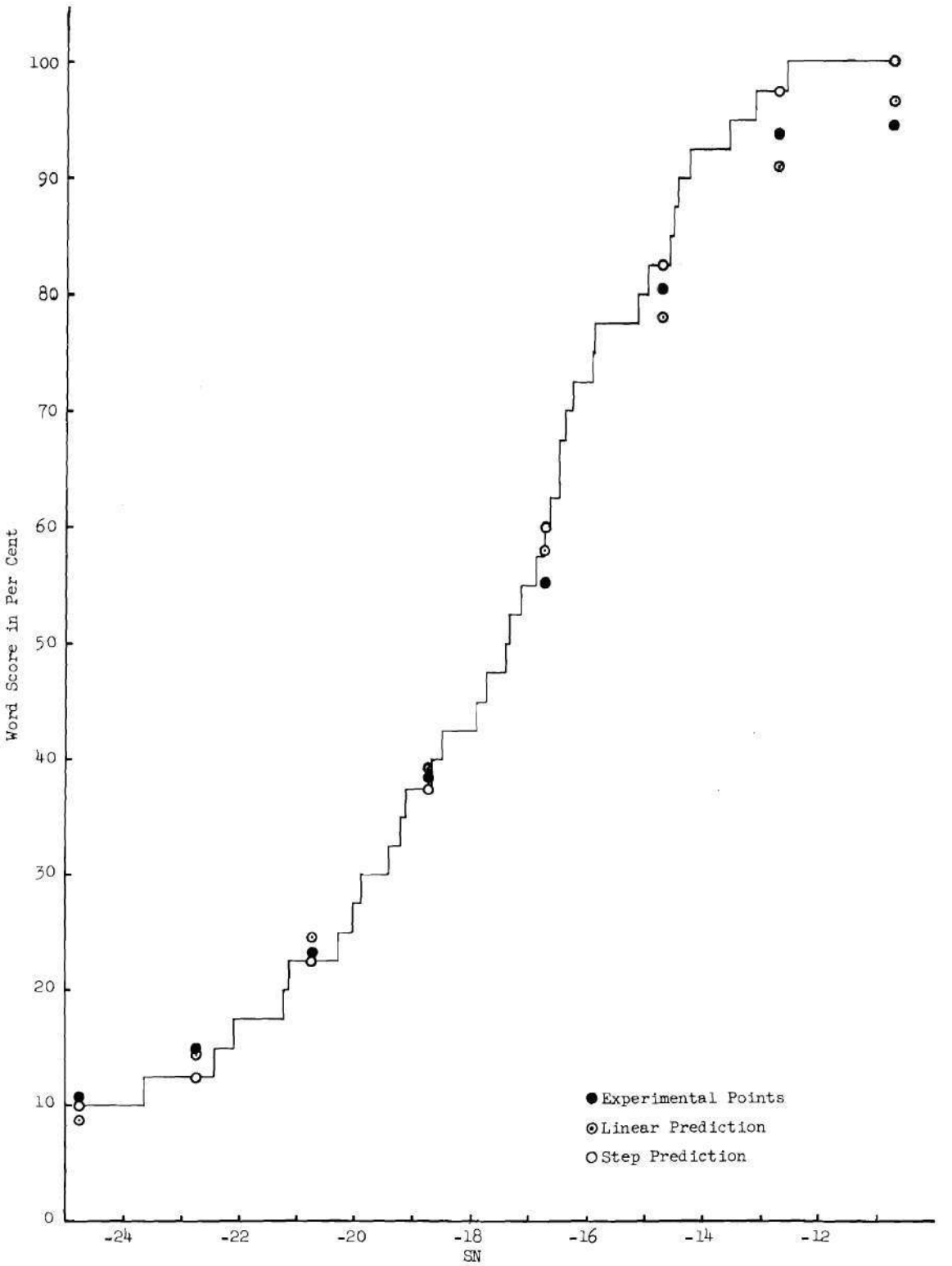


Figure 17. Experimental and Predicted Values of Word Score for JB (Master Set).

of score, while the maximum error for the linear approximation is only three percentage points. The largest difference between any two points is about nine percentage points between the linear and step prediction at SN = -12.75 db.

To facilitate a visual comparison, smooth curves were drawn through the sets of points of Figure 17, as shown in Figure 18. The two "predicted" curves show excellent agreement with the experimental curve and with each other over the entire test range. For example, the linear prediction and the experimental curve differ, at their 50 per cent levels, by only 0.25 db, while the step prediction differs from the experimental curve by only 0.3 db at this level. As one would expect from the nature of the approximation, the step prediction over-estimates the actual score at the high end of the curve, and under-estimates at the low end.

Following the same procedure, smooth curves were visually fitted to the three sets of points obtained by computing the actual curve (experimental scores), the linear prediction, and the step prediction, at each of the eight test values of SN and for each of the two remaining listeners. These curves are shown in Figures 19 and 20. Again, extremely good "prediction" is evident, except possibly in the case of the step prediction for listener NS. Finally, each of the curves was averaged over the listeners at each point to obtain "team curves" as shown in Figure 21. Figures 18 through 21 reveal uniformly good prediction from the linear approximation, with somewhat poorer, but still good, prediction from the step approximation. All of the curves exhibit an apparently typical tendency of the step prediction to over- and under-estimate at high and low scores, respectively, while at the same time agreeing closely

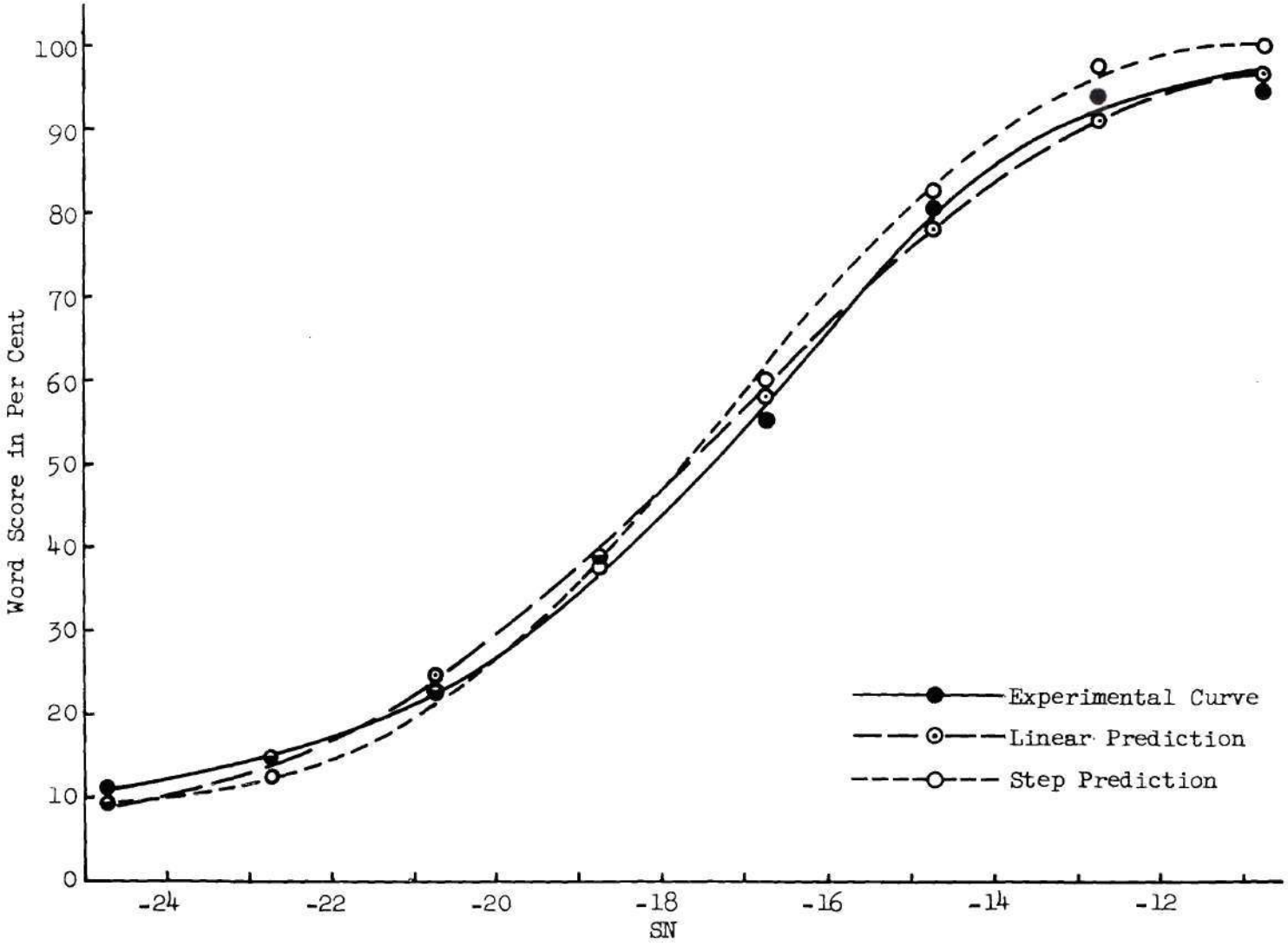


Figure 18. Experimental and Predicted Articulation Curves for JB (Master Set).

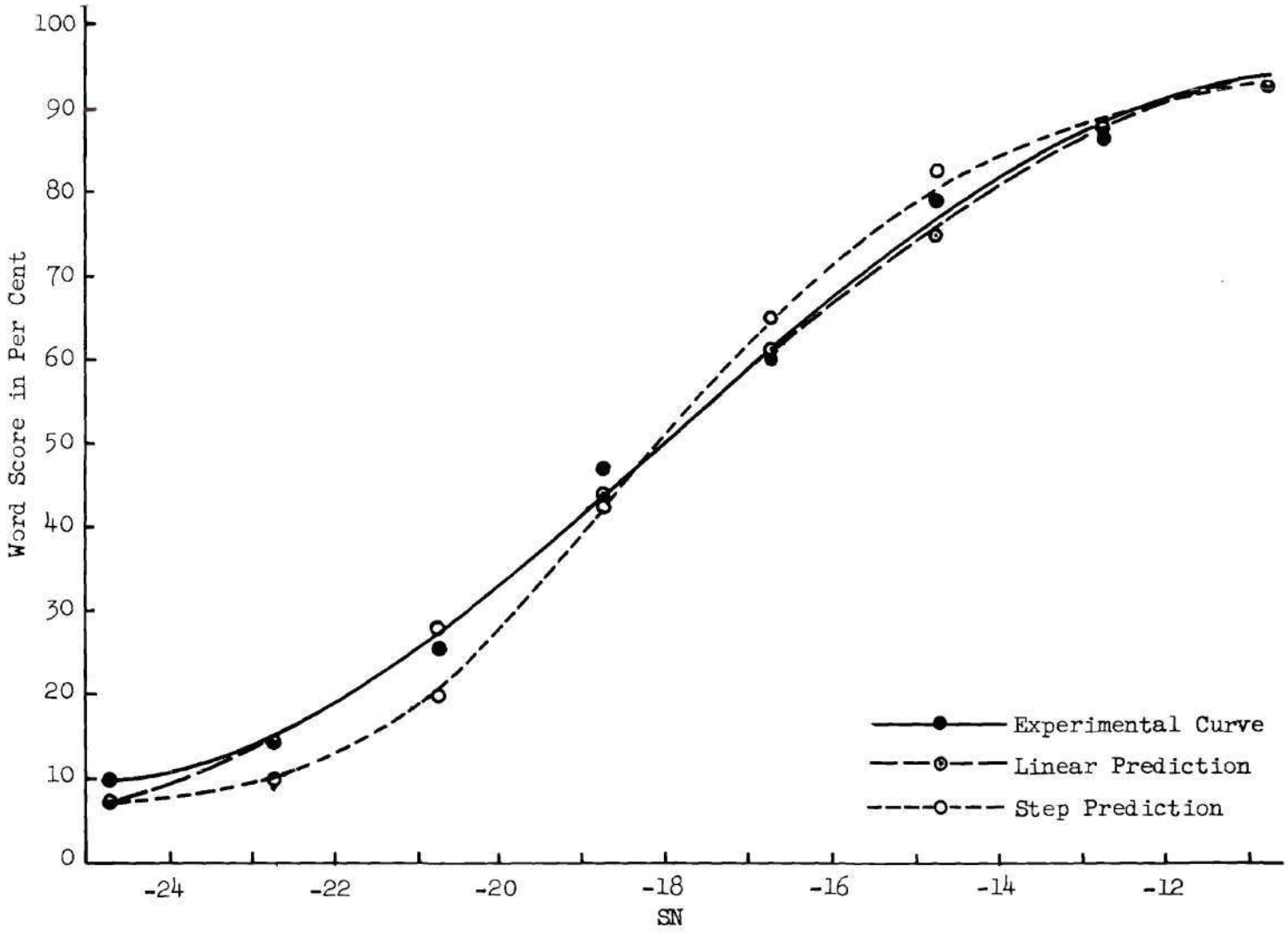


Figure 19. Experimental and Predicted Articulation Curves for WN (Master Set).

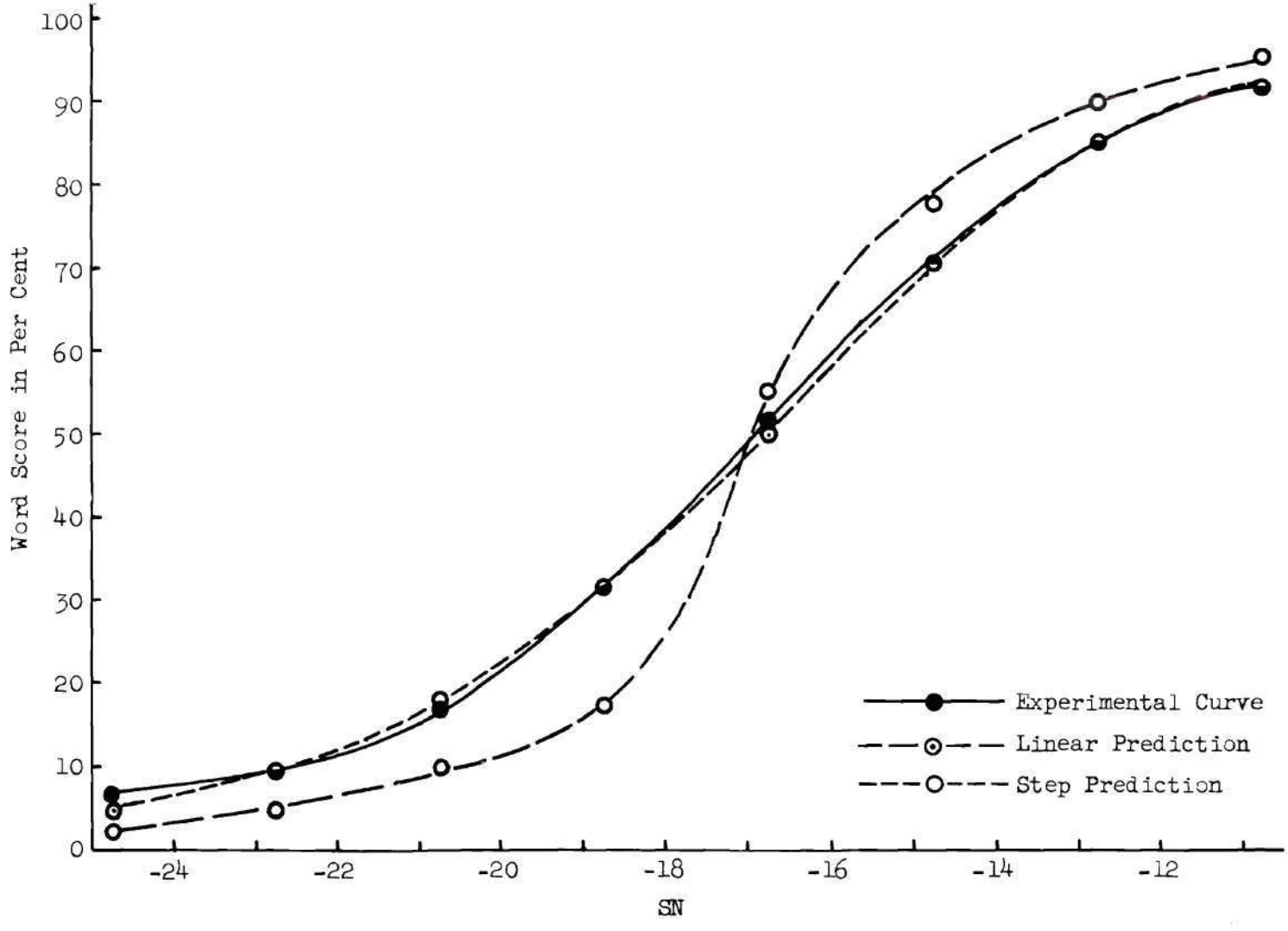


Figure 20. Experimental and Predicted Articulation Curves for NS (Master Set).

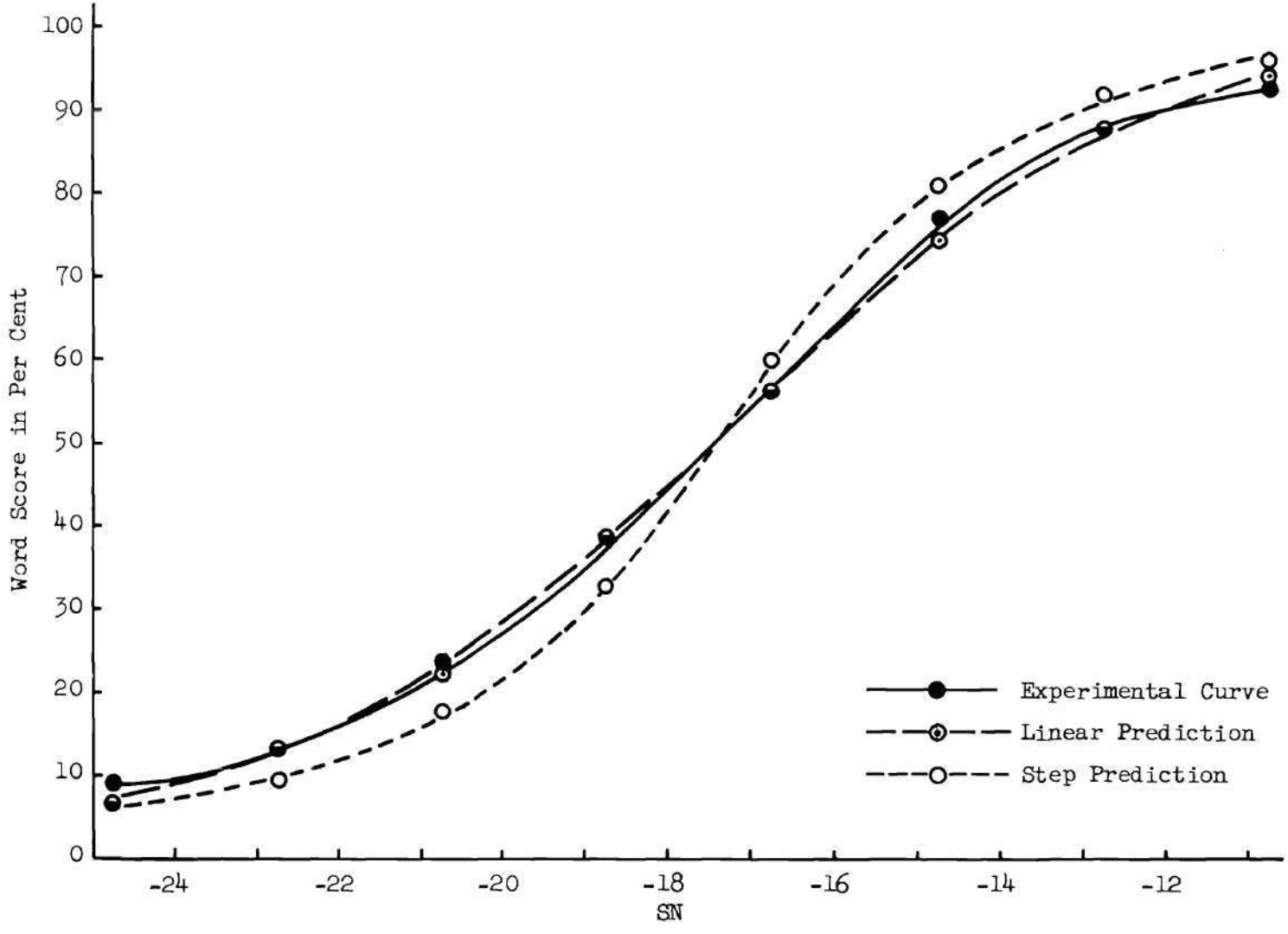


Figure 21. Experimental and Predicted Articulation Curves for Team (Master Set).

with the actual curve at the 50 per cent level.

To facilitate a comparison of errors, both here and for subset predictions described later, the following notation is used:

(a) ϵ_{ij} = directed distance in db along the SN coordinate, from curve i to curve j, at the 50 per cent level of score.

(b) η_{ij} = directed distance, in percentage points along the word score axis, from curve i to curve j, at the 50 per cent value of SN on curve i.

(c) Values of i and j range from 1 to 3, with 1 denoting the experimental curve, 2 denoting the linear prediction, and 3 denoting the step prediction.

Using the above terminology, the errors are tabulated in Table 5.

Table 5. Tabulation of Errors between Various Curves for Master Set

	JB	WN	NS	Team
ϵ_{12}	-0.25	0.00	+0.15	0.00
ϵ_{23}	-0.05	-0.10	-0.20	0.00
ϵ_{13}	-0.30	-.10	-0.05	0.00
η_{12}	+2.5	0.0	-1.5	0.00
η_{23}	+0.5	+1.0	+4.0	0.00
η_{21}	-3.0	0.0	+1.5	0.00
η_{31}	-3.5	-1.0	-0.5	0.00

Table 5 reveals extremely good agreement between the various curves at the 50 per cent level. The maximum difference magnitudes $|\epsilon_{12}|$,

$|\epsilon_{23}|$, $|\epsilon_{13}|$ are 0.25, 0.20, and 0.3 db, respectively, considering all three listeners, while the team curves show zero displacement. The average values (over the three listeners) of the difference magnitudes above are only 0.13, 0.12, and 0.15 db, respectively. The magnitudes $|\eta_{12}|$, $|\eta_{23}|$, $|\eta_{21}|$, and $|\eta_{31}|$, have maximum values over the three listeners of 2.5, 4.0, 3.0, and 3.5 percentage points of word score, respectively, with the team curves again showing perfect agreement. The average values over the listeners of these magnitudes are only 1.3, 1.8, 1.5, and 1.7 percentage points, respectively.

Based on the foregoing curves, there is justification for concluding, at least for test material of the general nature used in this study, that the two prediction schemes are sound and that the intelligibility characteristics of a word set are adequately described by the word parameters of threshold and NMGF spread. It follows that the shape of the articulation curve is dependent upon the distribution of these or of derived parameters, and hence that the curve may be shaped by choosing words on a basis of their threshold and spread. This is illustrated in following sections.

Subset Tests

Articulation tests on subsets of the master word-set had two purposes:

1. To evaluate the effectiveness of choosing subset words, on a basis of their threshold or spread, as a means of altering the location and spread of the articulation curve.

2. To evaluate the application of the prediction schemes to sub-

sets, where the word parameters used for prediction are obtained from tests on the master set.

When NMGF parameters obtained from master set tests are used to shape or predict articulation curves for subsets, the assumption made is that listeners respond to a given word, in its subset environment, in the same way that they responded to it in its master-set environment. This implies that when a listener is uncertain of a word, he chooses his response from the larger master-set, rather than from the smaller subset. For this to be true, the listener ideally should be unaware that he is listening to a reduced set, but this is difficult to achieve in practice unless words are repeated more than once. Such repetition is clearly undesirable if the master-set words were not so repeated, since the responses depend upon the relative frequency of occurrence of the words.

The next best approach would be to allow the listeners to realize that a smaller set is being used, but to prevent them, if possible, from determining which words are in the reduced set. This is essentially the approach followed here. Four 20-word subsets were chosen from the 40-word master set, and these were presented by transmitting each word only once during a given run, exactly as had been done for the master set. These subset tests were made under the same conditions and with the same general procedures as were used for previously-described tests. In order to prevent memorization of the subset vocabularies, several precautions were taken, as follows:

1. The number of repetitions of each subset was limited to five. (A repetition consisted of eight presentations of the subset at various

values of SN.)

2. The listeners were not given any training on the subsets. They did not know the words to be expected in a given subset, nor did they know which subset was being used.

3. The listeners were never told whether specific responses were correct or incorrect, nor were their scores available to them.

4. The listeners were required to cover up, on their score sheets, all previous responses, so that they would not be biased in their choices by having a record of previous responses. Score sheets were taken up immediately following each run.

5. The eight tapes available for each subset were in different random orders, and these were presented in such a way that no tape was presented twice at the same SN ratio, with the exception of one out of the total of 20 repetitions.

6. For two of the subsets, the word order on each tape was altered during the series of repetitions for that subset.

7. The subset tests were made in two parts. In the first half of the tests, two of the subsets were alternated randomly during the combined total of 10 repetitions. These repetitions, in turn, were interspersed with repetitions using "bogus" tapes, i.e., tapes containing the same number of words as the subsets being tested but having a different choice of words. The scores from these "bogus subsets" were not used, since their only purpose was to inhibit memorization of subset vocabularies. In effect, the listeners were presented with 10 different vocabularies rather than 2, during this half of the tests. The second half of the subset tests was made in exactly the same way, except that all "bogus" tapes were re-randomized.

8. The subjects were repeatedly reminded to avoid any conscious effort at memorizing vocabularies, and were told not to discuss the tests among themselves.

In spite of these precautions, some memorization apparently took place; this is discussed more fully in a later section. Also, there was probably a residual learning effect, but this was judged to be small. The effects of memorization, assuming an equal amount of memorizing on each subset, would be to alter the shape and location of each of the subset curves in approximately the same way. This clearly would affect the accuracy with which such curves could be predicted from master-set data, but should have much less effect on the relative shapes and/or locations of the curves. If, for example, two subsets were chosen on a basis of word thresholds, so as to give displaced articulation curves, then the curves should still be displaced after an equal amount of memorization on each subset. Only the absolute locations of the curves on the SN axis, and not their relative location, would be expected to change.

Shaping of Subset Curves

In order to illustrate the use of word parameters in shaping the articulation curve, it was desired to choose four subsets on the bases of word threshold and spread. Because of time limitations, it was necessary to fix the vocabularies for the subset tapes before data from the master-set tests could be completely processed, hence final values of β and Δ were not available for use in selecting the words. As a substitute, rough values of threshold and spread were computed, using data from five master-set repetitions and considering the three-man listening team as a single composite "listener." This was done as follows:

1. The data was examined for each listener at each of the eight SN values and each of the five repetitions. The scores (0 or 1) on a given transmission of one word were then tabulated separately for each word in the list, the tabulation covering all 40 transmissions of that word.

2. Scores were added for the three listeners, so that the "composite score" obtained on a given transmission was a number ranging from 0 to 3.

3. At each value of SN, these composite scores were summed over the five repetitions, thus yielding a sequence of eight numbers characterizing a given word. This sequence of numbers can be viewed, after division by 15, as ordered ordinates on a "composite NMGF" representing the intelligibility characteristics of that word, insofar as the team is concerned.

4. Each word's number sequence was examined for 50 per cent threshold and 13-to-87 per cent spread. The SN at which the sequence first increased to 2 or more (corresponding to $\frac{2}{15}$ x 100 or 13.3 per cent score) was recorded, as was the SN at which the sequence first reached or exceeded 13 (corresponding to 86.6 per cent score). Linear interpolation was used for sequences not containing the numbers "2" or "13." The difference in recorded values of SN was taken as the "spread" in db for that word. The value of SN at which the sequence first reached 7.5, (corresponding to 50 per cent score) was determined in a similar manner, and was taken as the "threshold" for that word.

The 40 words were divided into two subsets, denoted by "A" and "B," such that the 20 words having the lowest thresholds constituted subset A

and the 20 words having the highest thresholds constituted subset B. Subset A then consisted of the "most intelligible" words and subset B consisted of the "least intelligible" words.

The 40 words were again divided into two subsets, denoted by "G" and "H." Subset G consisted of the 20 words having the smallest spreads (largest NMGF slopes) and subset H consisted of the 20 words having the largest spreads (smallest NMGF slopes). The composition of these four subsets is given in Table 6.

Using the values of word power p^i and noise power p_n^i given in Table 2, the mean word and noise powers $\overline{p}$ and $\overline{p}_n$ and the group signal/noise ratio SN was calculated for each subset. For subsets A, B, G, and H, the values of SN (at the standard measurement point) were, respectively, -0.76 db, -0.73 db, -0.71 db, and -0.79 db. These are very close to one another and to the value of SN for the master set (-0.75 db), indicating homogeneity of all five sets with respect to SN.

The team scores resulting from the previously-described subset tests were tabulated, along with values of SN computed from equation (53), using the above values of group signal/noise ratio. The scores were then plotted versus SN, resulting in the articulation curves of Figures 22 and 23, where smooth curves have been visually fitted to the points.

As can be seen from Figure 22, the division of the master set on a basis of low or high threshold resulted in the well-displaced curves for subsets A and B. These curves are seen to have the same general shape and the same SN spread of about seven db between 20 per cent and 80 per cent score levels, but are displaced, at the 50 per cent level, by almost four db. Figure 23 illustrates the effects of choosing words on a basis

Table 6. Composition of Subsets

	Subset A	Subset B	Subset G	Subset H
1.	cast	ache	bald	ache
2.	crave	bald	cast	bead
3.	crime	bead	check	class
4.	deck	check	crave	dill
5.	dill	class	crime	fig
6.	fame	fig	deck	flush
7.	gnaw	flush	fame	gnaw
8.	hurl	gob	gob	leave
9.	muck	jam	hurl	lush
10.	neck	law	jam	nest
11.	pulse	leave	law	please
12.	rouse	lush	muck	pulse
13.	take	nest	neck	rate
14.	toil	path	path	take
15.	turf	please	rouse	thrash
16.	vow	rate	shout	turf
17.	wedge	shout	size	vow
18.	wharf	size	stag	wedge
19.	who	stag	toil	who
20.	why	thrash	wharf	why

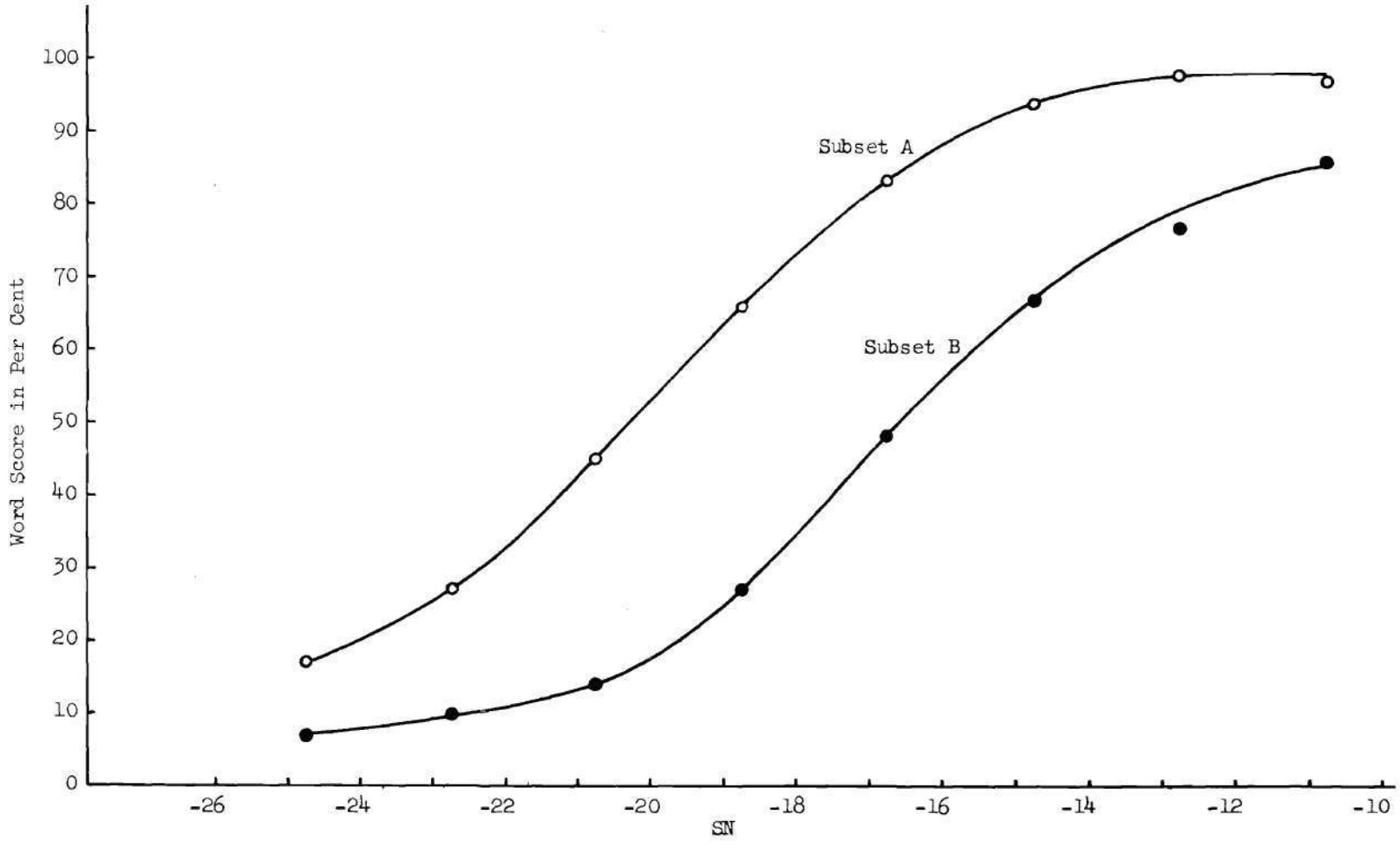


Figure 22. Experimental Articulation Curves for Subsets A and B.

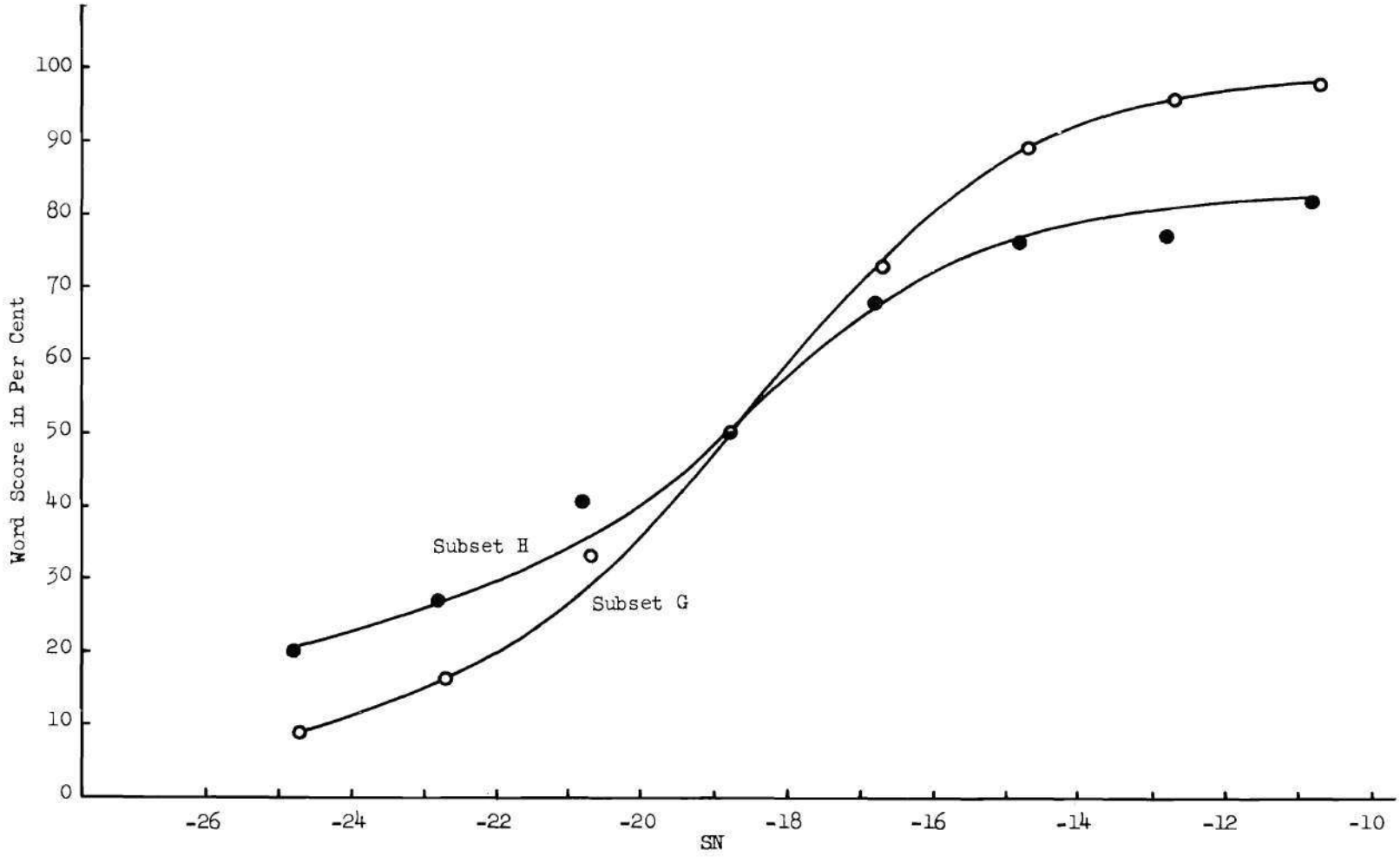


Figure 23. Experimental Articulation Curves for Subsets G and H.

of small or large NMGF spread. It shows that the low-spread set, subset G, and the high-spread set, subset H, have curves which are not displaced at the 50 per cent level but which have 20-to-80 per cent spreads that are quite different. These spreads are 6.0 db and 11.5 db for G and H, respectively, revealing a difference of 5.5 db in spread.

Apparently the shaping technique used, although somewhat crude, is quite effective. More sophisticated procedures, based on the prediction schemes and using values of β and Δ as defined in Chapter II, should be possible.

To illustrate the effects of the shaping technique just described, histograms were plotted for α and for Δ. Only valid values of these parameters (see Chapter III) were used in constructing these histograms, shown in Figure 24. The α-histograms are shown for subsets A and B and for each of the listeners, while Δ-histograms are shown for subsets G and H and for each of the listeners. The α-histograms for subsets A and B show clear separation of the centroid for all three listeners, illustrating the preponderence of low-threshold and high-threshold words, respectively. This effect is not quite so apparent in the Δ-histograms, although the curves for all three listeners indicate that the words having the largest spreads are contained in subset H. The valid values of α and Δ were next combined for the three listeners, and this data grouped by subset, resulting in the team histograms of Figure 25. The mean value is indicated by a dashed line on each curve.

Considering first the α-histograms, it is clear that good separation, in terms of threshold, was obtained for the team, in subsets A and B. The mean thresholds for subsets G and H are almost identical, although

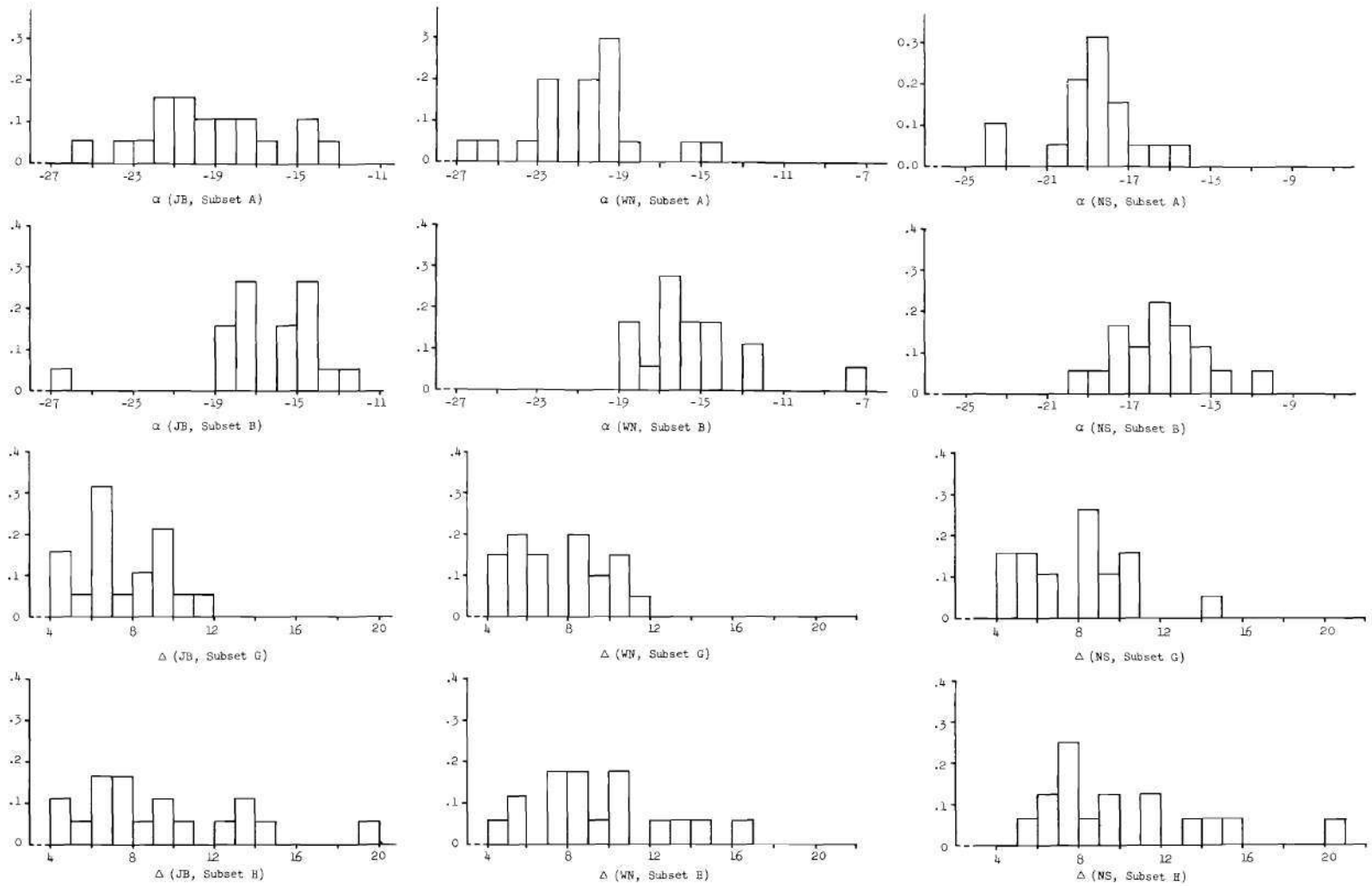


Figure 24. Histograms on α and Δ for Various Subsets and Listeners.

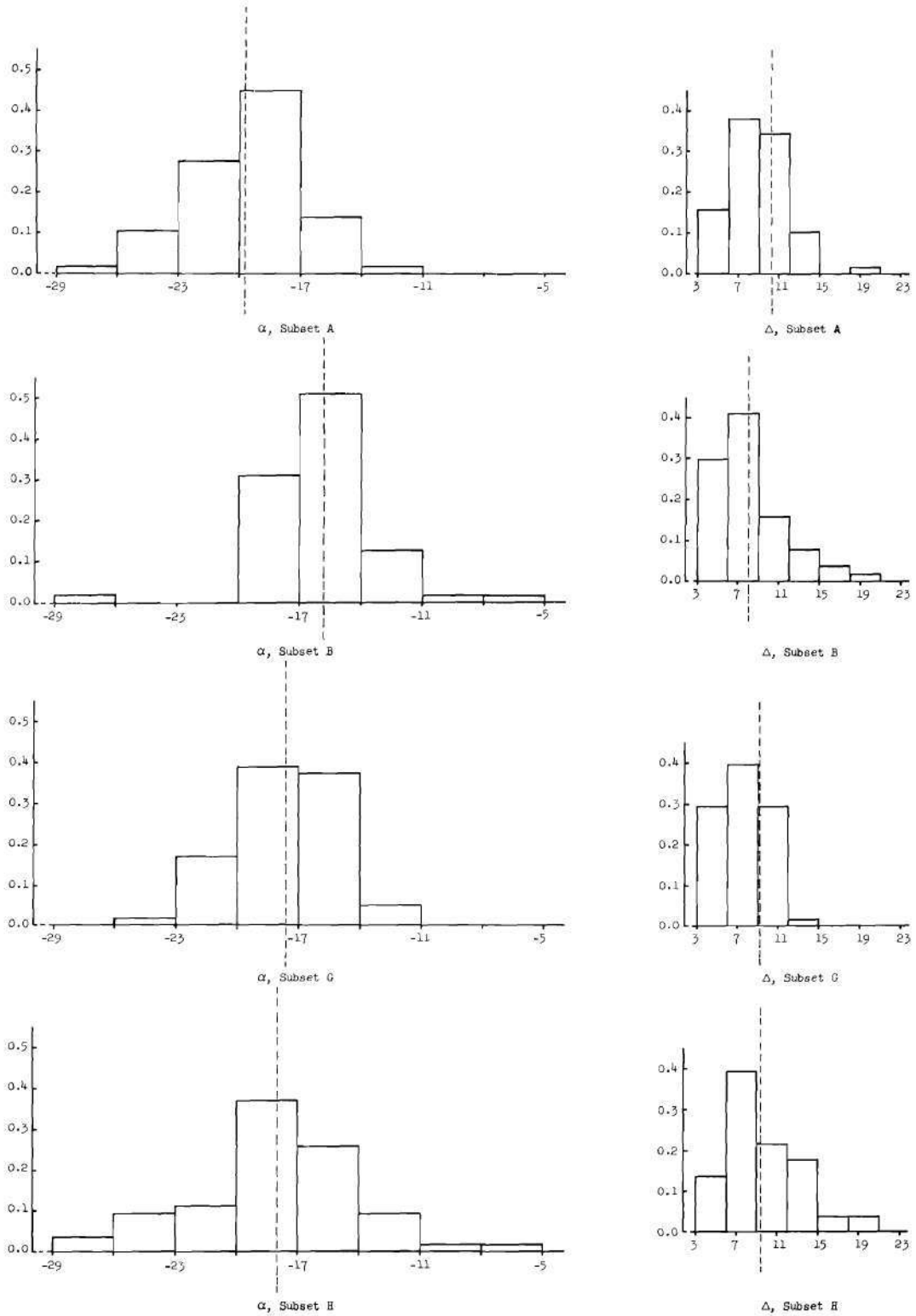


Figure 25. Histograms on α and Δ for Various Subsets (All Listeners).

most of the very high and very low thresholds are found in H. Considering the Δ-histograms, subsets A and B show considerable displacement, an unexpected result. Subsets G and H, chosen on a basis of small and large spread, show, contrary to expectations, very little displacement in mean spread, although most of the very large spread words are seen to be in H, as would be expected. Apparently, based on the curves of Figure 23, the inclusion of a small percentage of very large-spread words can significantly increase the spread of the articulation curve.

In summary, the threshold and spread parameters have been shown to be useful in shaping articulation curves. The possibilities in applying these ideas to the construction of curves having specified shapes are intriguing, especially in view of the fact that a third and important word parameter, namely, relative power, is also available for use in a sophisticated shaping scheme.

Prediction of Subset Curves

At this point, the prediction schemes have been evaluated only for application to the master set, i.e., to the same set used for determining the word parameters. The real test of these schemes as a means of predicting articulation curves arises in applying them to subsets.

Following the procedures described earlier, the step-approximation and linear-approximation schemes were applied to predict articulation curves for subsets A, B, G, and H, using values of β and Δ determined from tests on the master set and values of SN given for the four subsets in the preceding section. Note that values of β must be recalculated for each different subset, since by equation (54) this parameter depends on the SN at standard level, and the latter quantity is different

for different subsets.

Points on the predicted curves were calculated for each subset and each listener, but only at the test values of SN. This was done to facilitate comparison with the actual scores obtained from the subset tests, these being available only at the test points. Points on the step prediction, linear prediction, and experimental curves were plotted on the same graph for each listener-subset pair, and smooth curves visually fitted to the points. These are shown in Figures 26 through 33, each set of curves being identified as to subset and listener. Figures 32 and 33 illustrate curves obtained by averaging the actual and predicted curves over the three listeners and are therefore labeled as "team curves."

The figures show generally good agreement between the experimental curves and the predicted curves. As was the case for the master set, the step predictions tend to over- and under-estimate the linear prediction at the high and low ends, respectively. The linear prediction generally agrees more closely in shape with the actual curve than does the step prediction, but in nearly all cases a distinct displacement of the linear prediction from the actual curve is seen to occur. This displacement is in the direction of higher scores for the actual subset tests than are predicted from master-set data and is most pronounced near the low end and in the center region of the curves. Despite this displacement, which is discussed in the following section, the predicted curves are satisfactory estimates of the actual curves for many purposes. The average error between the experimental curve and the linear prediction, at the 50 per cent level, is only about 1.3 db for the 12 individual listener

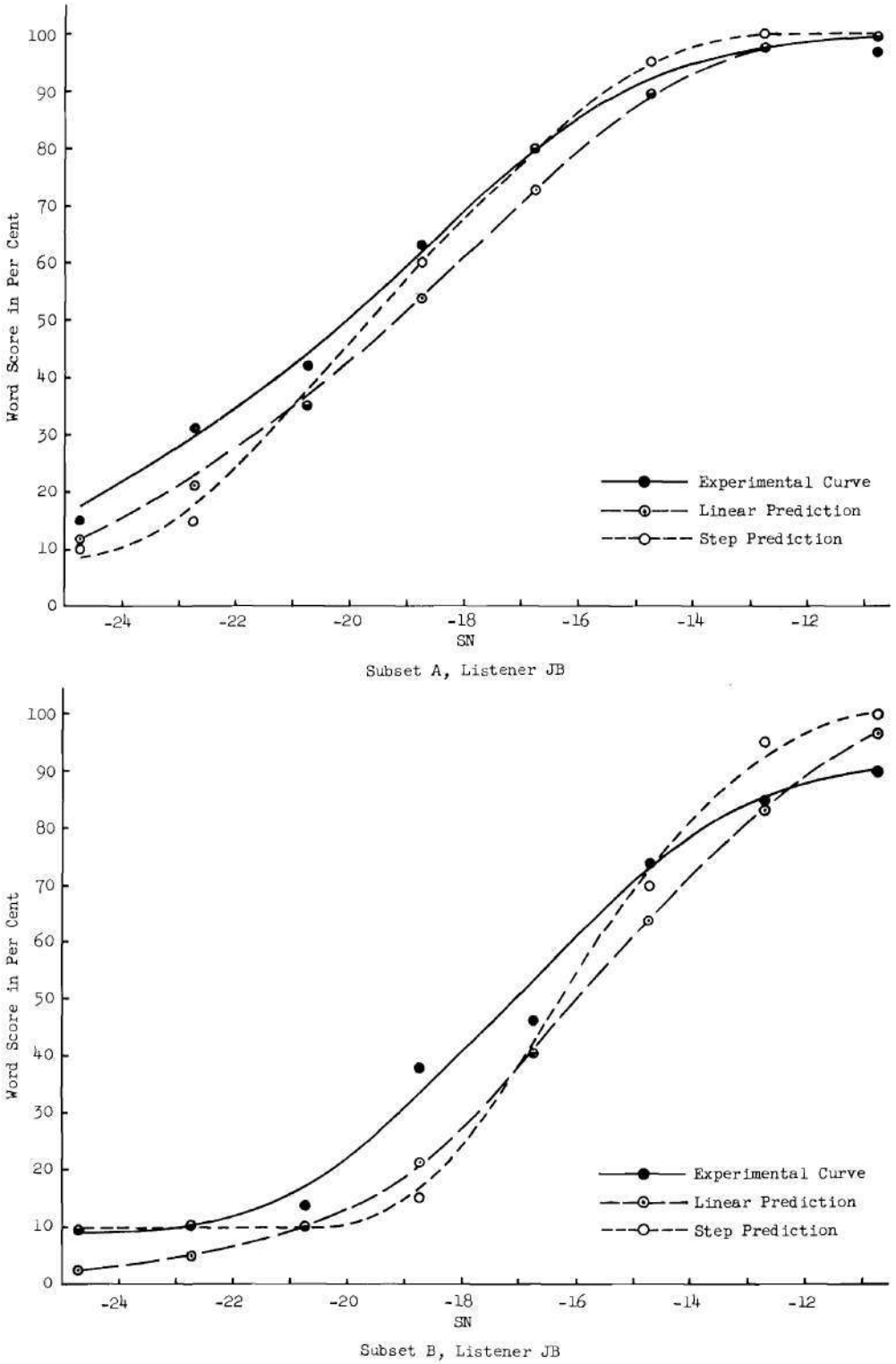


Figure 26. Experimental and Predicted Curves for Subsets A and B, Listener JB.

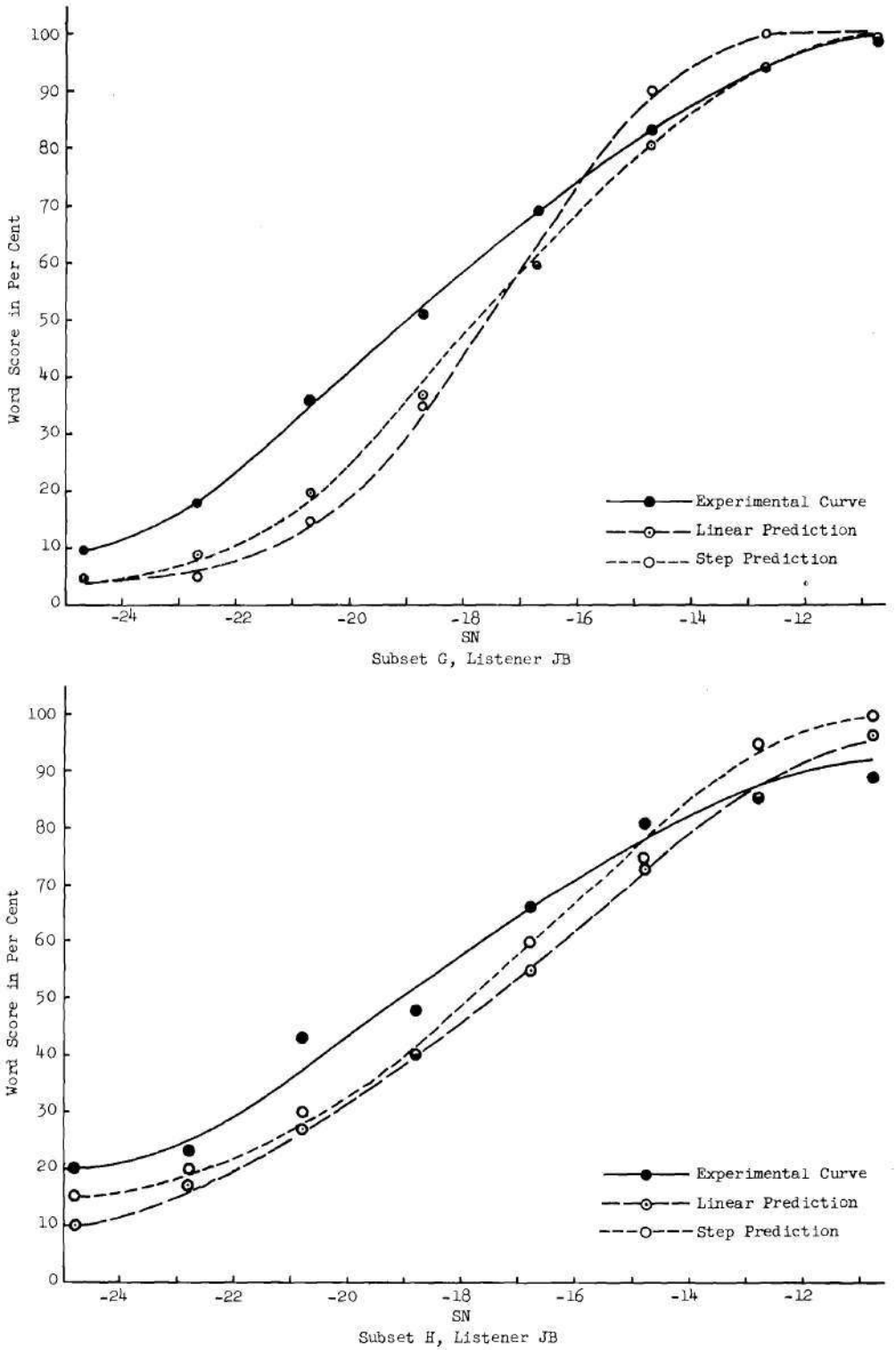


Figure 27. Experimental and Predicted Curves for Subsets G and H, Listener JB.

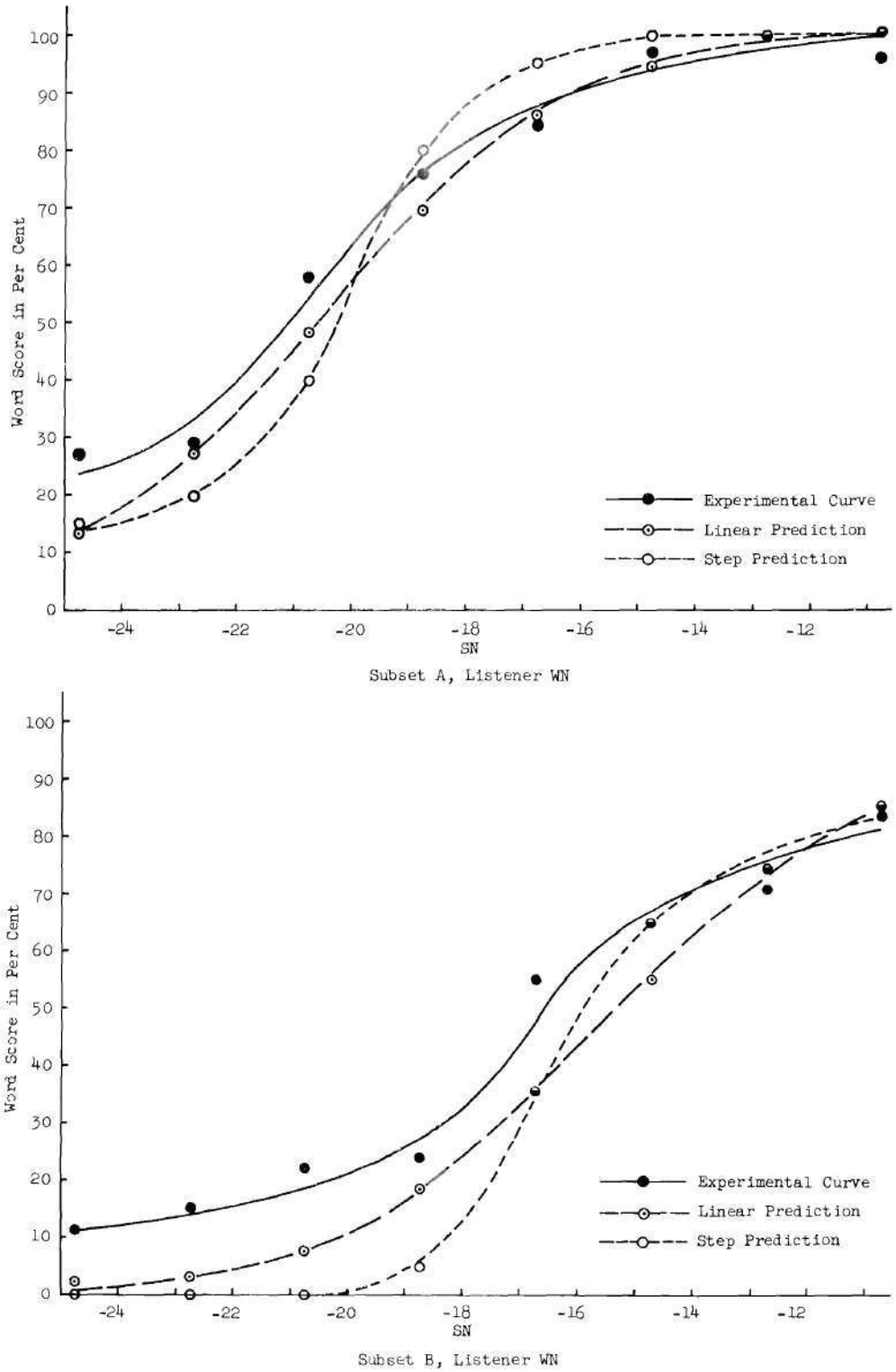


Figure 28. Experimental and Predicted Curves for Subsets A and B, Listener WN.

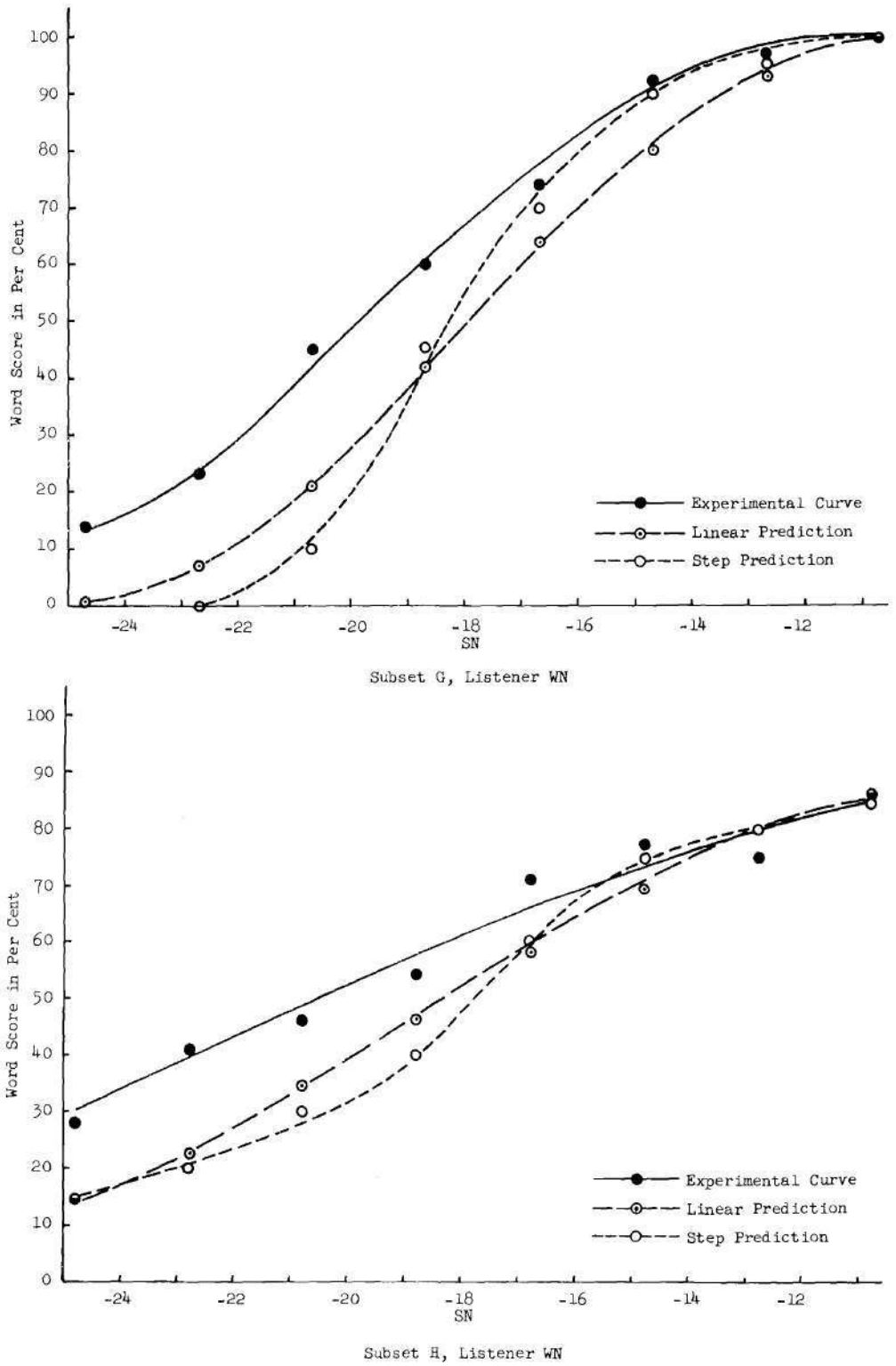


Figure 29. Experimental and Predicted Curves for Subsets G and H, Listener WN.

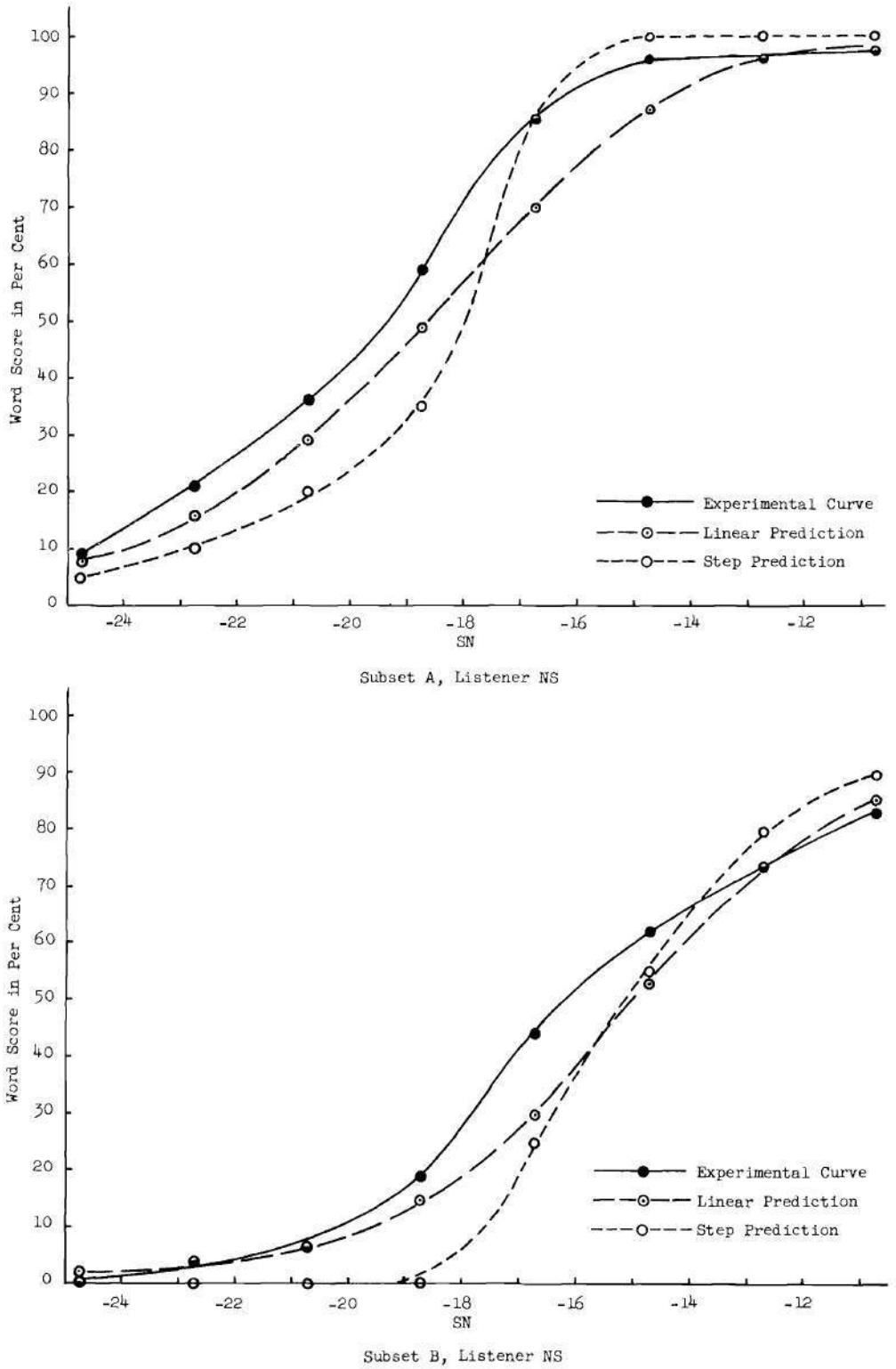


Figure 30. Experimental and Predicted Curves for Subsets A and B, Listener NS.

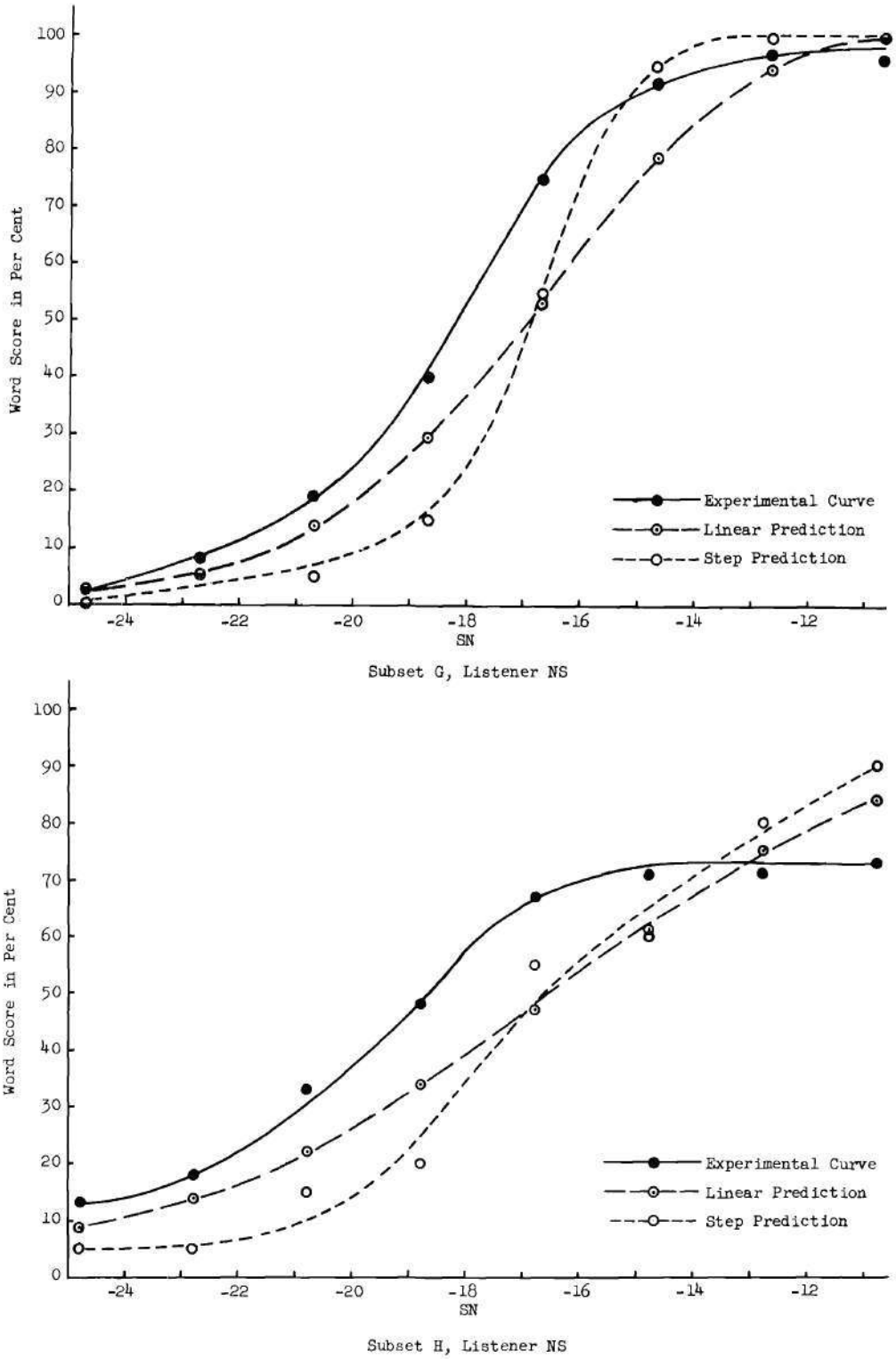


Figure 31. Experimental and Predicted Curves for Subsets G and H, Listener NS.

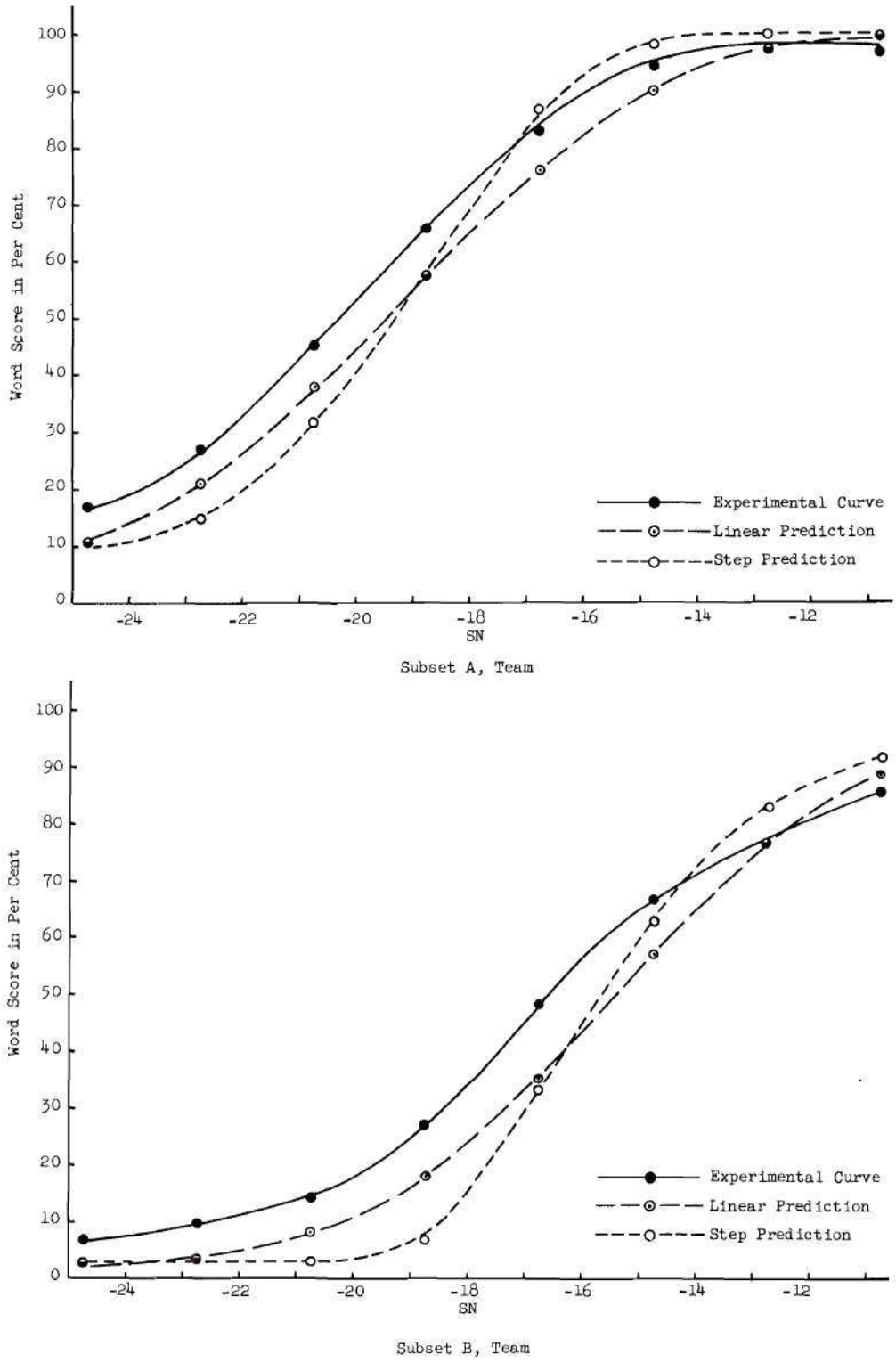


Figure 32. Experimental and Predicted Curves for Subsets A and B, Listening Team.

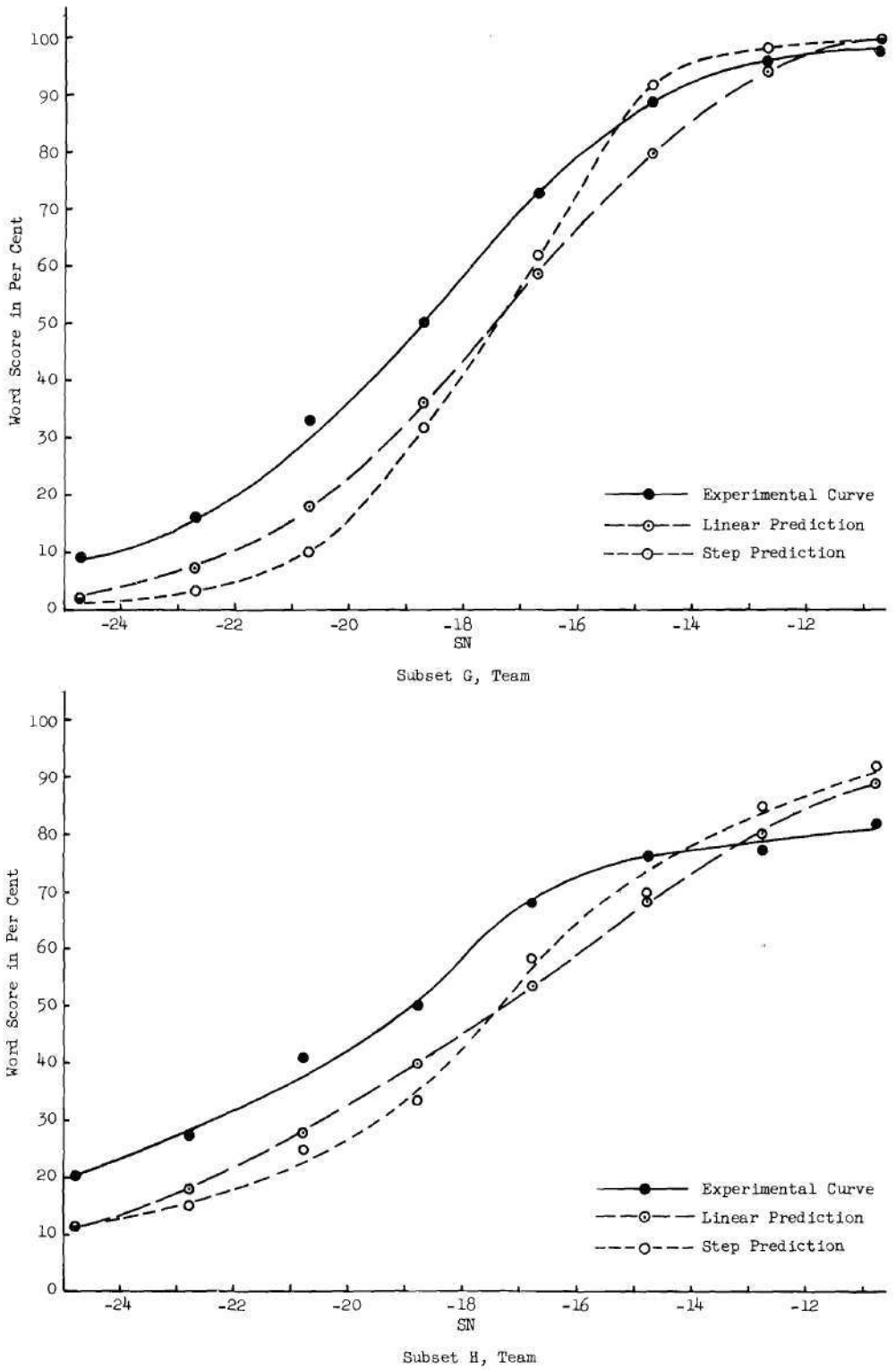


Figure 33. Experimental and Predicted Curves for Subsets G and H, Listening Team.

curves and is only 1.2 db for the 4 team curves. Such accuracy is quite satisfactory in situations where the location of the 50 per cent level is needed to within only two or three db, or in cases where the intelligibility is being estimated at a value of SN which is known only to this accuracy. Such situations occur, for example, in the analysis of military communication system performance.

For a more detailed analysis of discrepancies between the various curves at the 50 per cent level, the quantities ϵ_{ij} and η_{ij}, as defined in the first part of this chapter, were obtained. As before, the possible subscript values are 1, 2, and 3, denoting experimental, linear prediction and step prediction curves respectively. The values of these quantities are distances between curves, in db for ϵ and percentage points of score for η. The magnitudes of these errors are tabulated in Table 7. From the data in this table, maximum and average values (over the three listeners) were obtained for the error magnitudes. These, along with values of error magnitude for the team curves, are summarized in Table 8.

The average and team values for ϵ and η in Table 8 give some idea of what error magnitudes to expect in using the prediction schemes, at least for tests involving roughly the same conditions as those described here. First, the error between the linear and step approximation, as given by $|\epsilon_{23}|$, is generally very small. In spite of the over- and under-estimation of the step prediction (with respect to the linear one) which occurs at high and low values of score, the two curves agree very closely at the 50 per cent level, the maximum error observed being only 0.7 db for all subsets and listeners, and the "grand average" error over all listeners and subsets being only 0.34 db. The error for the team

Table 7. DB and Score Errors between Various Curves at the 50 Per Cent Level

	Subset:	A	B	G	H
$\lvert\epsilon_{12}\rvert$	JB	0.85	1.05	1.25	1.60
	WN	0.50	1.20	1.95	2.15
	NS	0.70	1.20	1.25	2.10
$\lvert\epsilon_{23}\rvert$	JB	0.45	0.30	0.15	0.45
	WN	0.35	0.50	0.30	0.55
	NS	0.70	0.15	0.05	0.10
$\lvert\epsilon_{13}\rvert$	JB	0.40	0.75	1.40	1.15
	WN	0.85	0.70	1.65	2.70
	NS	1.40	1.05	1.30	2.00
$\lvert\eta_{12}\rvert$	JB	8.0	12.5	14.0	12.0
	WN	6.0	12.5	21.5	14.0
	NS	7.5	14.0	15.5	15.5
$\lvert\eta_{23}\rvert$	JB	5.0	4.5	3.0	3.5
	WN	7.5	7.5	6.0	6.0
	NS	13.0	2.0	2.0	1.0
$\lvert\eta_{21}\rvert$	JB	7.5	11.0	10.5	11.5
	WN	6.0	12.5	17.5	9.5
	NS	10.5	10.0	21.0	18.0
$\lvert\eta_{31}\rvert$	JB	3.5	7.5	11.5	8.5
	WN	10.0	8.5	15.0	11.5
	NS	22.0	8.5	22.0	17.5

Table 8. Maximum, Average (Over Three Listeners), and Team Values of Error Magnitudes at the 50 Per Cent Level

Type of Error	Subset A	Subset B	Subset G	Subset H	Average Over Subsets
$\lvert\epsilon_{12}\rvert_{max}$	0.85	1.20	1.95	2.15	1.54
$\lvert\epsilon_{23}\rvert_{max}$	0.70	0.50	0.30	0.55	0.51
$\lvert\epsilon_{13}\rvert_{max}$	1.40	1.05	1.65	2.70	1.70
$\lvert\epsilon_{12}\rvert_{av}$	0.68	1.15	1.48	1.95	1.32
$\lvert\epsilon_{23}\rvert_{av}$	0.50	0.32	0.17	0.37	0.34
$\lvert\epsilon_{13}\rvert_{av}$	0.88	0.83	1.44	1.95	1.28
$\lvert\epsilon_{12}\rvert_{team}$	0.85	1.20	1.25	1.50	1.20
$\lvert\epsilon_{23}\rvert_{team}$	0.15	0.25	0.05	0.10	0.14
$\lvert\epsilon_{13}\rvert_{team}$	1.00	0.95	1.30	1.40	1.16
$\lvert\eta_{12}\rvert_{max}$	8.0	14.0	21.5	15.5	14.8
$\lvert\eta_{23}\rvert_{max}$	13.0	7.5	6.0	6.0	8.1
$\lvert\eta_{21}\rvert_{max}$	10.5	12.5	21.0	18.0	15.5
$\lvert\eta_{31}\rvert_{max}$	22.0	8.5	22.0	17.5	17.5
$\lvert\eta_{12}\rvert_{av}$	7.2	13.0	17.0	13.8	12.8
$\lvert\eta_{23}\rvert_{av}$	8.5	4.7	3.7	3.5	5.1
$\lvert\eta_{21}\rvert_{av}$	8.0	11.2	16.3	13.0	12.1
$\lvert\eta_{31}\rvert_{av}$	11.8	8.2	16.2	12.5	12.2
$\lvert\eta_{12}\rvert_{team}$	8.5	13.0	14.0	10.0	11.4
$\lvert\eta_{23}\rvert_{team}$	2.0	3.5	1.0	1.0	1.9
$\lvert\eta_{21}\rvert_{team}$	9.0	12.0	15.0	15.0	12.8
$\lvert\eta_{31}\rvert_{team}$	10.5	10.0	15.5	14.0	12.5

curves is even smaller, a relationship which can be observed to hold for most of the tabulated values.

Although the error magnitudes do not add (i.e. $|\epsilon_{13}|$ is not equal to $|\epsilon_{12}| + |\epsilon_{23}|$), one would generally expect, because of the relatively small values of $|\epsilon_{23}|$, about the same error between the actual curve and either of the predictions. This is borne out by the tabulation in Table 8, where the maximum observed prediction error is seen to be 2.15 db for $|\epsilon_{12}|$ and 2.70 db for $|\epsilon_{13}|$. That is, if one predicted the 50 per cent value of SN by either of the schemes, the maximum error to be expected during tests on several subsets with several listeners should be on the order of 2.5 db. Likewise, the average prediction error to be expected, considering all listeners and subsets, should be on the order of 1.3 db. Again, the prediction error for team curves is somewhat smaller, its average value over all subsets and listeners being about 1.18 db for either prediction scheme.

The percentage score errors, as given by η, are numerically much larger than the db errors; this is due to the large slope (typically about 10 percentage points per db) of the curves near the 50 per cent level. Except for this numerical factor of roughly 10, the η's follow essentially the same pattern as the ϵ's, that is, the linear and step predictions agree fairly closely at the 50 per cent level, while yielding roughly the same error when used to predict the actual scores. Again, the score errors for the team curves are never much greater, and in most cases less, than the corresponding errors for individual curves. Finally, when prediction errors averaged over listeners are compared with those averaged over subsets, the latter are found to be both larger and more

variable than the former, indicating that choice of subset has somewhat more to do with the accuracy of prediction than choice of listener.

In summary, the prediction errors resulting from either scheme are reasonably small, for many purposes, at the 50 per cent level of intelligibility. A conservative estimate of the maximum and average error to be expected, under conditions similar to those described previously, is 2.0 and 1.5 db, respectively.

The slopes at the 50 per cent level, in percentage word score per db of SN, were calculated for each listener, subset, and type of curve, using the curves of Figures 18 through 21, and Figures 26 through 33. These, along with averages over the three listeners, are given in Table 9. Slopes for the master-set curves are included for comparison. As can be seen from the table, the average slopes for subsets A and B are very nearly the same for each of the three types of curves, whereas the slopes for G and H are quite different, as would be expected. The step prediction slopes are uniformly higher than corresponding ones for the linear prediction and actual curves, whereas the latter two have somewhat similar slopes. Finally, the slope for the actual curve, averaged over all listeners and all subsets, is 10.5 per cent per db, compared to 9.9 for the master set and an average of 10.3 for the team subset curves. These results compare closely with corresponding figures for the linear prediction.

Discussion of Results

The fact that master-set curves are uniformly "predicted" with great accuracy by the linear approximation scheme indicates not only that this scheme accounts in a valid way for the contributions of individual

Table 9. Slopes at 50 Per Cent Level, in Per Cent per DB, for Various Curves

Curve	Word Set	Listener				
		JB	WN	NS	Average Over Listeners	Team
Step Prediction	Master	12.0	12.0	25.0	16.3	14.3
	A	11.0	20.0	27.3	19.4	13.9
	B	16.4	18.7	15.0	16.7	15.4
	G	14.0	19.0	30.0	21.0	16.1
	H	9.3	10.2	10.0	9.8	11.6
Linear Prediction	Master	10.0	8.8	11.1	10.0	9.3
	A	9.1	12.0	11.1	10.7	11.8
	B	12.3	10.0	12.0	11.4	11.3
	G	11.7	10.7	13.8	11.8	12.0
	H	7.7	6.7	7.0	7.1	6.8
Actual	Master	10.7	8.8	10.3	9.9	9.7
	A	8.6	12.3	13.0	11.3	10.5
	B	10.4	15.0	10.0	11.8	11.3
	G	8.8	9.5	16.7	11.7	11.4
	H	7.0	4.6	9.8	7.2	7.8

words to the articulation curve, but also that the set of NMGF thresholds and spreads contain essentially all of the information concerning the intelligibility of a word set. These conclusions are supported by 40-word curves for individual listeners as well as for the team; intuitively the conclusions should be even more valid when larger numbers of

words are used. It seems reasonable to expect qualitatively similar results on speech units other than monosyllables, for types of masking other than white noise, and possibly for channel parameters other than signal/noise ratio.

The curves and tabulated errors indicate that the step-approximation scheme is as accurate as the linear one, at intelligibility levels near 0, 50, and 100 per cent, while diverging somewhat at intermediate points. Although only a limited number of listeners was used, it is of interest that for all three listeners, as well as for the team, the slope of the actual master-set curves can be obtained in a consistent way from the step-predictions for this set. Specifically, points are located on the step prediction at four db above and below its 50 per cent point, and a straight line is drawn between the points. The slope of this line closely approximates that of the actual curve, which, in turn, is fairly approximated, between its 40 and 60 per cent levels, by such a line. This same consistent estimation of actual curve slope from the step prediction can also be made for subsets, if only the team curves are considered. The displacement of actual from linear or step predictions in the case of subsets is a different matter, such displacement, as explained later, being dependent on the degree to which subset responses are chosen from the master-set vocabulary. In the present case, where such choices were not made consistently, the resulting displacement for team curves is still remarkably uniform from subset to subset, being 1.0, 1.0, 1.3, and 1.5 db for subsets A, B, G, and H respectively. The foregoing facts imply that the step prediction can be used to obtain a fair estimate of the actual curve over most of its range, provided, in the case of subsets,

that there is either no memorization of subset vocabulary or else a known "memorization shift" for at least one subset of each size.

A comparison of the team subset curves has been made to illustrate the possibility of shaping the articulation curve by choosing words on a basis of threshold and spread. The two subsets chosen on the former basis have curves which were displaced, one to the left and one to the right of the master-set curve. The subsets chosen on a spread basis exhibited curves differing in spread and, as would be expected, in slope. The slopes, at the 50 per cent level, were 11.4 and 7.8 per cent per db for subsets G and H, respectively (team curves). As was pointed out earlier, more sophisticated shaping techniques should be possible, utilizing not only choice of words but adjustment of individual word power. It is interesting to note from Table 9 that the various team curve slopes for subsets are roughly twice as great as that reported by Curry and others (22) for the 26 alphabet letters while for the master set team curve the slope is also about twice the value obtained by others for 50 PB words. The slightly smaller number of stimulus items used for the present study explains part of this difference in each case but such comparisons are difficult, at best, due to differences in speaker, number of listeners, and listener training.

The main error in predicting subset curves was a fairly consistent displacement of predicted and actual curves which was more pronounced at low and medium values of SN. Although the prediction accuracy is sufficiently good for many purposes, some effort was made to determine the source of this error. It has already been postulated that the main source was partial memorization of the subset vocabulary; this apparently

occurred in spite of the precautions taken. Such memorization obviously would improve the scores above the values predicted on a basis of no memorization. A rough qualitative explanation, given below, of why scores would be increased more at the low end of the curve than at the high end, involves the number of guesses made by the listeners in responding to the transmitted words.

It is not possible, from score sheet data alone, to determine which responses were guesses (since some guesses turn out to be correct), but it is possible to determine the number of wrong guesses, simply by counting the number of word responses which are incorrect. It should be pointed out here that some responses were "blanks" (nothing heard); these are not counted as guesses. At small SN, the scores are generally small and the number of wrong guesses approximates the total number of guesses, assuming, as was the case, relatively few "blanks." The data was examined at the two lowest test values of SN, and the average (over listeners, subsets, and values of SN) number of wrong guesses determined to be 14.8 out of the 20 responses. This indicates that at least 74 per cent of the responses near the low end of the curve were guesses, as compared to only a few per cent near the high end.

If partial subset memorization occurs, then the responses would tend to be words from the partially known subset vocabulary, rather than from the master set. In the extreme case of complete memorization (amounting to training the listeners on subset words), each guess would be a choice selected from the 20 subset words rather than the 40 master set words, and the probability of guessing correctly, on a given trial, would be doubled. In any event, one would expect a larger number of

correct guesses than if no memorization occurred, purely on the basis of choosing from a smaller set. At the same time, the fraction of total correct responses attributable to correct guesses is larger at the low end of the curve where relatively more guessing is done. Hence any event (such as memorization) which makes correct guesses more likely will affect scores at the low end more than at the high end. In the present case, one would expect to find the improvement in lower scores to be larger, percentage-wise, than for higher scores, when comparing the actual curve to the predicted one. This is the general effect observed in Figures 26 through 33. Before proceeding to formulate this explanation in mathematical terms, it is pertinent to observe that in the extreme case of training on subsets, i.e., complete memorization of vocabulary, some experimental curves have been published by Miller (16) which can be used to approximate an upper bound on the shift produced by changing the size of the word set. Miller's curves are articulation curves for 2, 4, 8, 16, 32, 256, and 1000 monosyllables, and show a progressive shift to the left (smaller SN for a given score) as the number of test items is reduced. The shift can be estimated for vocabulary sizes other than the above by plotting, at a given score, the SN versus number of words, using Miller's curves. When this is done, on semi-log paper, and a smooth curve fitted to the points, the SN resulting in the given score can be read from this curve for any number of test items. This procedure was followed, using a score value of 50 per cent and the resulting displacement of 20- and 40-word curves determined to be about 1.7 db. The actually observed shift in team curves (prediction error), averaged over the four subsets, is about 1.2 db, indicating something less than complete memorization.

That the observed shift in curves was real, and not due to some consistent calculation error or defect in the linear prediction scheme, was verified by plotting subset curves directly from the appropriate set of NMGF's, using actual NMGF points rather than a linear approximation. The set of subset curve points so obtained fitted the linear prediction curve extremely well in every case, except at the lowest and highest point on each curve, and even at these points the discrepancy was fairly small. A quantitative explanation of the effect is thus desirable; this can be developed by assuming a simplified model of the stimulus-response mechanism for a listener.

In this simplified model, the following assumptions are made:

(1) The N-word vocabulary is known in every case, and the listener makes a forced-choice response to each transmitted word.

(2) When a word is transmitted, the listener either understands it unambiguously (absolute certainty) or else has no clue whatever and makes a pure guess. This is equivalent to assuming a step-like NMGF for all words.

(3) For pure guesses, the response is not biased by any previous response and is chosen with equal likelihood from the available N words. This is equivalent to assuming that guesses are independent Bernoulli trials with probability $p = \frac{1}{N}$ of "success" (correct guess) on each trial.

(4) The fraction of total words accounted for by unambiguous ones depends on, and only on, SN, this dependence being fixed and independent of N. This is equivalent to assuming that the distribution of thresholds F_β is independent of N, since this fraction is the fraction of words above threshold, and hence is given by $F_\beta(SN)$. For convenience here,

this quantity is denoted simply by F, a monotone non-decreasing function of SN ranging in value from 0 to 1.0.

The N responses to each test consist of a certain number n of pure guesses and a number N-n of unambiguous responses. Then

$$\frac{N-n}{N} = F , \tag{55}$$

and hence

$$n = N(1 - F) . \tag{56}$$

The fractional score on a given test is then

$$\text{Score} = \frac{1}{N} (\text{total correct responses}) \tag{57}$$

$$= \frac{1}{N} (NF + S_n) ,$$

where S_n is the number of correct guesses out of the n total guesses, i.e., the number of successes in n Bernoulli trials. S_n is thus a binomially-distributed random variable, with a probability that exactly k successes will result given by (39, chapter 9)

$$\text{Prob}\left[\text{k successes in n trials}\right] = b(k; n,p) \tag{58}$$

$$= \frac{n!}{k!\,(n-k)!}\, p^k (1-p)^{n-k} .$$

The expected value of the score is, from (57),

$$E\left\{\text{Score}\right\} = F + \frac{1}{N} E\left\{S_n\right\} = F + \frac{np}{N} \tag{59}$$

Using equation (56) and the fact that $p = \frac{1}{N}$, this becomes

$$E\left\{\text{Score}\right\} = F + \frac{1}{N}(1 - F) , \qquad (60)$$

where the first term represents the contribution to the score of the unambiguous words and the last term represents the contribution of correct guesses. Assuming a specific function for F, the score varies with SN as shown by the solid curve in Figure 34.

To better illustrate the effect of N, equation (60) is rearranged as follows.

$$\text{Score} = \frac{1}{N} + F\left(1 - \frac{1}{N}\right) \qquad (61)$$

Referring to equation (61) and Figure 34, it is seen that, for values of SN less than SN_o, F is zero and the fractional score is constant at $\frac{1}{N}$. For values of SN greater than SN_1, F is unity and the score is constant at 1.0. As SN increases from SN_o to SN_1, the score rises monotonically according to equation (61), describing the solid-line articulation curve.

The effect of transmitting a subset of N' words chosen from the original N words is now investigated for two extreme conditions:

(a) The listener receives no training on the new N'-word vocabulary (no memorization of subset vocabulary) and still makes guesses as before from the original N-word vocabulary.

(b) The listener is re-trained on the N'-word vocabulary and makes all guesses from this reduced set.

In most articulation testing, condition (b) is maintained. In

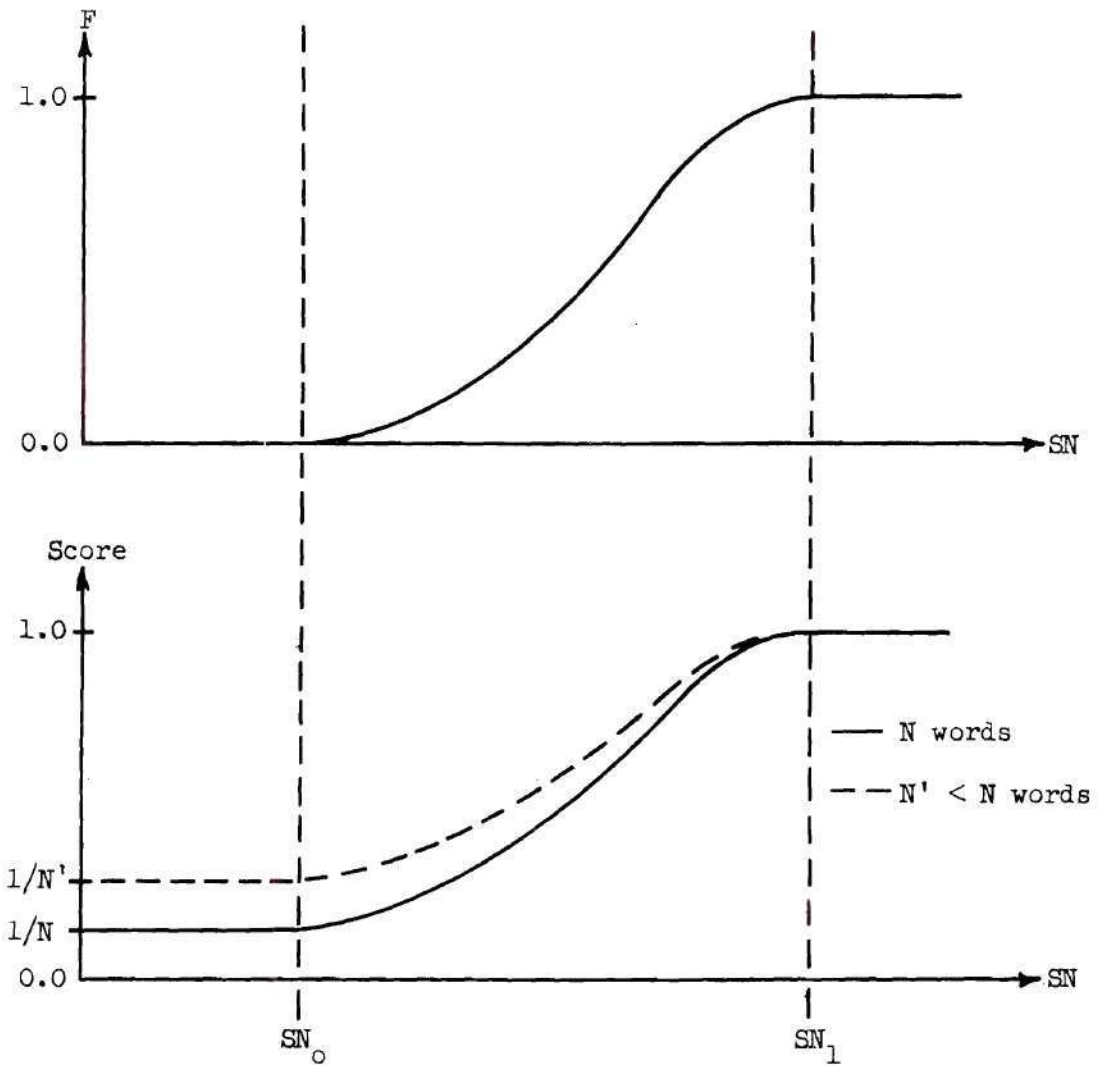


Figure 34. Variation of Score with SN.

the tests described here, an attempt was made to maintain condition (a), but the actual result was somewhere between (a) and (b), i.e., partial memorization. The prediction schemes are, of course, based on (a), and hence exhibit discrepancies in the present case. First, it will be shown that under (a) there should be no discrepancy, and then it will be shown that under (b) the discrepancy is of the same nature as actually occurred.

Under condition (a), the probability p remains at its original value $\frac{1}{N}$, but equation (57) becomes

$$\text{Score} = \frac{1}{N'} (N'F + SN), \tag{62}$$

and the expected value of the score becomes

$$E\left\{\text{Score}\right\} = F + \frac{np}{N'}, \tag{63}$$

where, from equation (56), $n = N'(1-F)$ and $p = \frac{1}{N}$. Hence the expected score is

$$E\left\{\text{Score}\right\} = F + \frac{1}{N}(1 - F) \tag{64}$$

$$= \frac{1}{N} + F\left(1 - \frac{1}{N}\right)$$

which is identical with equation (61). Hence, under the assumptions on F, the articulation curve is unchanged and is still represented by the solid-line curve of Figure 34.

Now under condition (b), the probability p becomes $\frac{1}{N'}$ and, as before, the expected score becomes

$$E\left\{\text{Score}\right\} = \frac{1}{N'} + F\left(1 - \frac{1}{N'}\right) . \qquad (65)$$

When equation (65) is plotted, assuming the same F as before, the result is the dashed curve in Figure 34. This curve is identical with the one for condition (a) and $SN \geq SN_1$, but is constant at a new value $\frac{1}{N'} > \frac{1}{N}$ for $SN \leq SN_o$. Hence the discrepancy between the two curves is greatest near the low end of the curve, and gradually decreases as the scores increase. Near the low end of a curve for N words, the value of F is small, and, from equation (61), the determining factor is N. Near the high end of the curve, F is nearly unity and the determining factor is F. If, as assumed, F is the same for different values of N, then the high end of the curve is relatively unaffected by a reduction of N to N', while the low end is greatly affected by such a change. At intermediate values of SN, the discrepancy is less than at SN_o, since the F factor then accounts for a larger percentage of the score, and the "guessing factor" $\frac{1}{N}$ accounts for a smaller percentage of the score, than was the case near SN_o.

In summary, the guessing is more important in determining score the low end of the curve and less important at the high end. Hence any factor (such as memorization of the subset vocabulary) which increases the probability of a correct guess will raise the low end of the curve more than the high end. This general effect is observed in comparing the actual and linear prediction curves of Figures 26 through 33.

The non-validity of the four initial assumptions made in this simplified model will affect the above conclusions, as follows:

(1) If the response is not forced-choice, the presence of "blank"

responses will make guessing relatively less important (since fewer guesses are made), and tend to reduce the predicted discrepancy. The discrepancy should still be of the nature predicted by the simple model, however, and will be present as long as any guesses are made.

(2) If some guesses are "informed guesses" and not random, the scores will tend to be higher at all points on the curve, both for condition (a) and condition (b). As long as any random guesses are made, however, a discrepancy, of the predicted nature, should remain. The more clear-cut the division of words into unambiguous and completely ambiguous, i.e., the more nearly step-like the NMGF's, then the better the accuracy of the simplified model.

(3) If guesses are not random, but are biased by previous responses, the resulting "bias effect" is difficult to evaluate. However, this effect is intuitively of the same order of magnitude, as far as score is concerned, for both N and N' words, assuming memorization, and hence should not qualitatively alter the predicted discrepancy.

(4) If F (which is actually the threshold distribution) is not the same for N and N', then the simplified model is clearly not valid. To the extent that the set of thresholds of the N' subset words have the same distribution (or histogram) as that of the N words, the model is valid.

From the above, it is seen that the nature of the discrepancy predicted by the model is not essentially changed provided only that the last assumption (invariant F) is approximately true. Thus the observed discrepancy, being of the nature predicted by the model, can reasonably be assumed to result from the increased probability of correct

guesses which accompanies memorization of the subset. It is of interest to compare, for different subsets and listeners, the degree to which the predicted discrepancy actually occurred, and to correlate this with the degree to which various subsets and listeners fit the simplified model.

First, although not so instructed, the listeners approximated a forced-choice type of response, in that "blanks" were relatively infrequent, as compared to wrong guesses. Near the low end of the curve, where the number of correct responses is small, the number of wrong guesses is approximately equal to the total number of guesses (such guesses have already been stated to comprise 74 per cent of all responses at the two lowest values of SN). At the lowest SN, the ratio λ was calculated for all listeners, subsets, and repetitions, where

$$\lambda = \frac{(\text{number of wrong responses})-(\text{number of blanks})}{(\text{number of wrong responses})} \qquad (66)$$

$$= \frac{(\text{number of wrong guesses})}{(\text{number of wrong responses})}$$

$$\doteq \frac{(\text{number of forced-choice guesses})}{(\text{number of ambiguous words})} .$$

When averaged over all listeners and subsets, λ was found to be 0.886, indicating roughly that 87 per cent of the ambiguous words had been responded to by a forced choice, and only 13 per cent had been responded to by blanks. Hence assumption (1) is approximately correct. Further, when examined by listener (averages over subsets) and by subset (averages over listeners) it was found that λ had the following average values:

JB:	0.90	WN:	1.00	NS:	0.76		
A:	0.85	B:	0.83	G:	0.95	H:	0.91

Referring to Figures 26 through 33, it is seen that the nature of the discrepancy agrees most closely with that predicted by the model for listener WN, as would be expected because of his perfect compliance with the forced-choice assumption. Also, the subset agreeing most closely with the predicted effect is subset G, as would be expected from the fact that the listeners tended to make more forced-choice responses for that subset than for any other. In particular, the discrepancy for subset G and listener WN illustrates almost perfectly the effect predicted by the model, λ for this case being 1.00. The main anomalies are for listener NS, who responded with a much larger percentage of blanks than either JB or WN.

Secondly, the subset having the most nearly step-like NMGF's (assumption (2)) is subset G, in fact this was the basis of choice for words in this subset. Again, this subset exhibits the predicted discrepancy more clearly than any of the others.

Finally, the two sets having threshold histograms most nearly like those of the master set (Assumption (4)) are G and H, from Figures 25 and 43. Again, the agreement with the predicted nature of the discrepancy is more marked for these subsets than for A or B.

Based on the foregoing discussion, it seems reasonable to accept the simplified listening model as an explanation of discrepancies between predicted and actual subset curves. Some comments are in order concerning the assumption of no training on the subset. Although the prediction schemes make this assumption, it should be possible to modify the schemes, by assuming a more complex model of the listening process, so as to take into account various amounts of memorization. The most important practical

case is that of complete memorization (re-training on subsets); this is the situation in many articulation tests. A practical situation analogous to the test conditions and assumption of no training on subsets would be a communication system restricted to a relatively small vocabulary and through which messages (subsets of the master vocabulary) are sent. The messages should be constrained to have either a very low redundancy or else roughly equal redundancy, so that guesses for ambiguous words are either random or made with roughly equal probabilities of success. Such a situation could conceivably arise in military or control tower voice communication.

All of the results of applying the prediction schemes, both to the master set and to subsets, indicate that the set of thresholds plays a leading role in determining the shape of the articulation curve. To the extent that this curve is obtainable from the step prediction scheme, one can state that the shape is determined by F_β, the distribution of SN thresholds (see equation (26)). Another way of putting this is that β, given by

$$\beta = \alpha - P + P_n \ , \tag{67}$$

is the basic factor in shaping the curve. From previous definitions, however, β is expressible as

$$\beta^i = 10 \log \left[\frac{p^i}{p_n^i} \right]_t - 10 \log \frac{p^i}{\bar{p}} + P_n^i \tag{68}$$

$$= 10 \log \left[\frac{\left[\frac{p^i}{p_n^i} \right]_t}{\frac{p^i}{\bar{p}}} \right] + P_n^i$$

$$= 10 \log \gamma^i + P_n^i ,$$

where $\left[\frac{p^i}{p_n^i} \right]_t$ is the threshold value of word signal/noise ratio, expressed as a numeric, and γ^i is the ratio of this threshold to the normalized word power, also expressed as a numeric. Thus, for a given set of noise powers, i.e., for given values of P_n^i, the basic determinant of the shape of the articulation curve is seen to be γ, the ratio of word threshold to word power. In the case where the P_n^i are equal, i = 1,2,....,N, the distribution of β depends only on the distribution of γ, and one can state that the shape of the articulation curve depends on the distribution of threshold-to-power ratios of the individual words.

CHAPTER V

SUBSIDIARY RESULTS

In this chapter are presented some miscellaneous results which, although not necessarily unimportant, are subsidiary to the major objectives of the research. For the most part, these results are presented with only brief comment.

Energy, Duration, and Power Distributions

The measurement of various physical parameters associated with the speech and noise waveforms was discussed in Chapter III. From the measured values it is possible to plot histograms and distribution functions which give some idea of how these parameters are distributed in magnitude. Caution is indicated in extending any conclusions drawn from such curves to the general case of monosyllabic words, primarily because of the relatively small amount of data available. In particular, the 40 measured values of each parameter are not sufficient, in plotting the histograms, to provide a good approximation to the probability density function of the underlying random variable. This can be seen by observing that the choice of different intervals in plotting the histograms causes significant changes in the shape of the curve. For the most part, fairly large intervals were used in order to present the gross characteristics of the distribution.

In Figure 35 is shown a histogram for the numerical values of word energy w. As in following histograms, the ordinate in each interval is

the fraction of observed values falling within that interval. The histogram indicates that low values of energy predominate. The total spread in energy covers a range of 4.52 db, and the average energy, indicated by the dotted line, is about 35.6.

In Figure 36 is shown a histogram of word duration T with the abscissa in milliseconds. The average duration (547 milliseconds) is indicated by a dotted line. This curve indicates that values of T are fairly uniformly distributed over the master set. When the percentage spread (total spread in values divided by the mean value) is calculated for w and T, the values are, respectively, 122 and 39.6 per cent, indicating that the value of T is much more constant, relatively speaking, than is the value of w. Hence it is not surprising that values of the ratio $p = w/T$ (word power) tend to be distributed somewhat like w, i.e., low values predominate. This is illustrated in the word power histogram of Figure 37, where a dotted line has been drawn through the mean word power ($\overline{p} = 64.9$).

When word power in db relative to $\overline{p}$, i.e., P, is considered, the resulting histogram appears as shown in Figure 38. The total range of word power is 3.77 db, approximately 0.77 db of which is accounted for by the most powerful word ("gob").

Cumulative distribution functions F_w, F_T, F_p, and F_P are shown in Figures 39, 40, 41, and 42 respectively. These curves give a more accurate and detailed picture of the way in which the word energy, duration, and power are distributed than do the corresponding histograms. The energy distribution function reveals the concentration of values in the low-energy region, while the distribution of durations, as expected

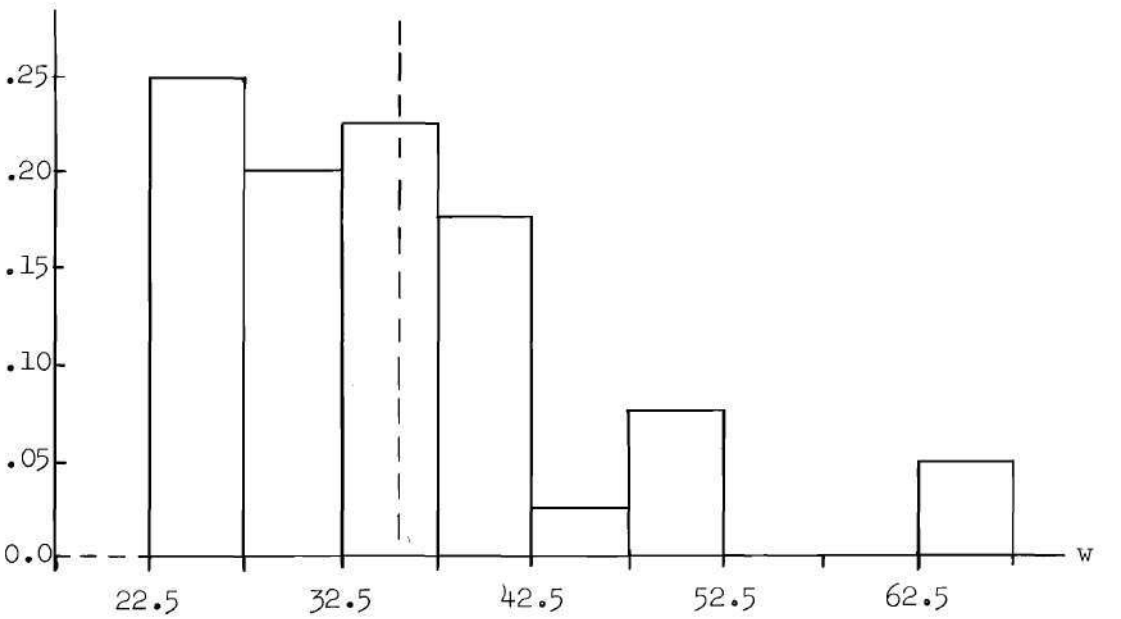


Figure 35. Histogram for w, Master Set.

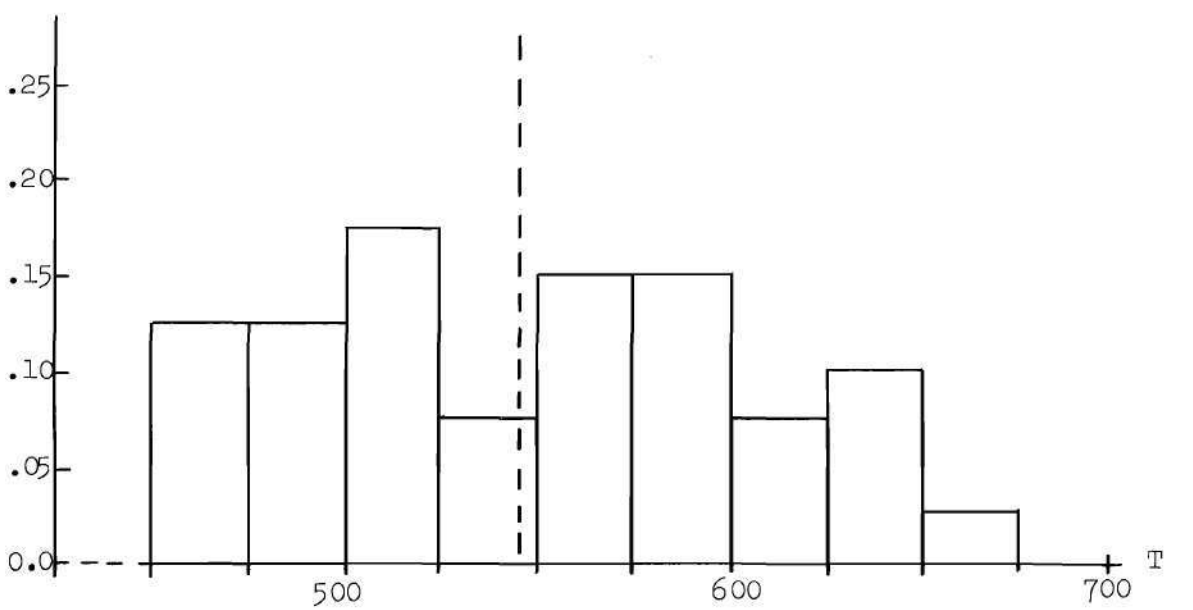


Figure 36. Histogram for T in Milliseconds, Master Set.

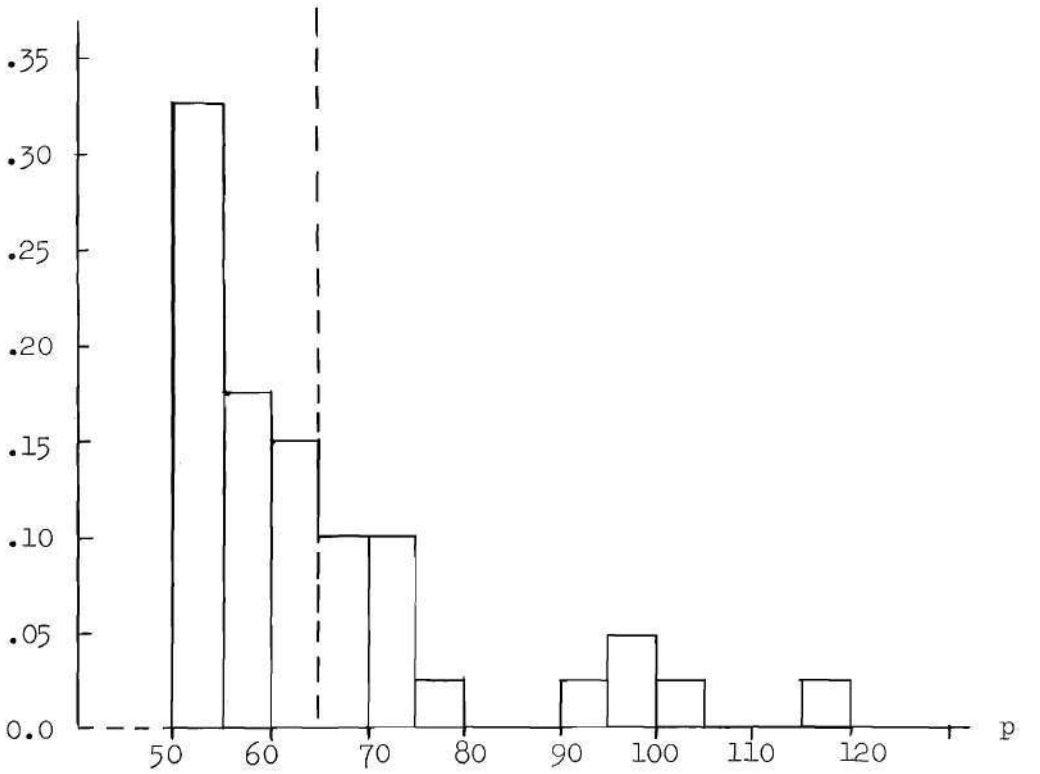


Figure 37. Histogram for Word Power p, Master Set.

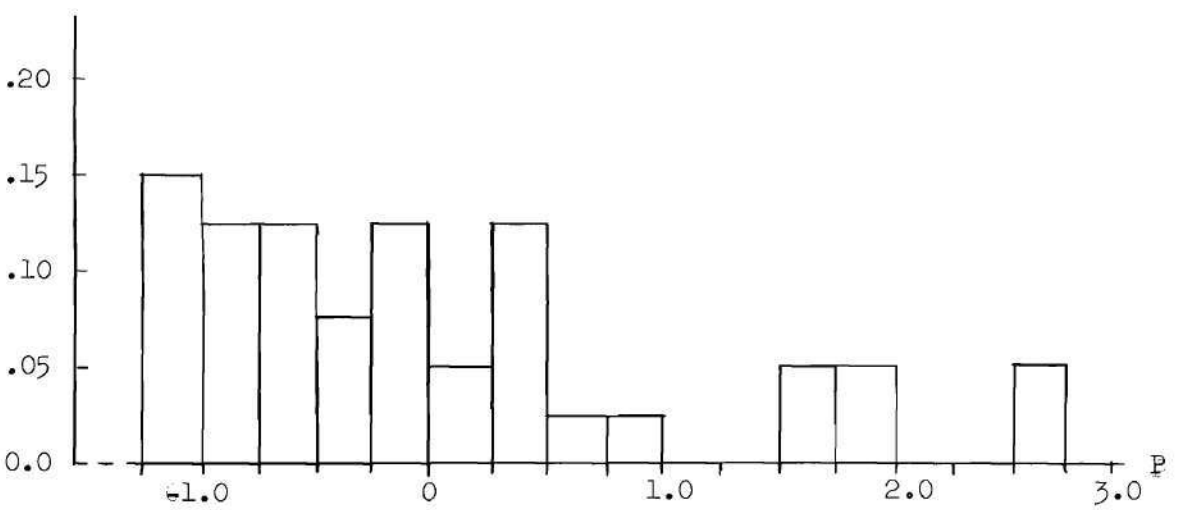


Figure 38. Histogram for Word Power P in DB, Master Set.

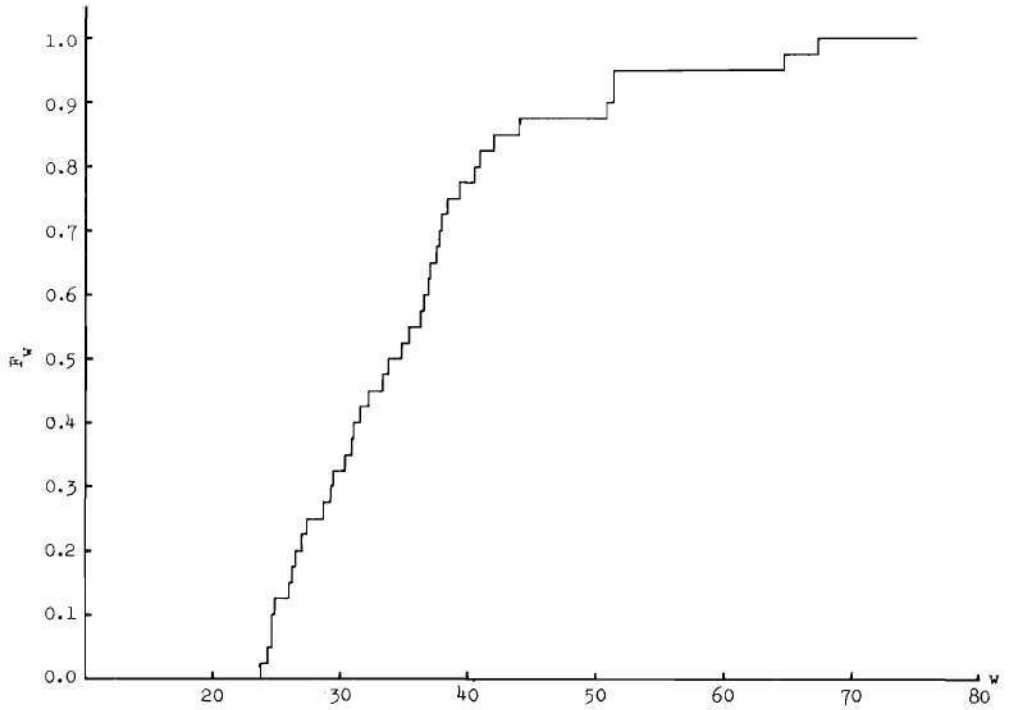


Figure 39. Distribution Function for Word Energy w, Master Set.

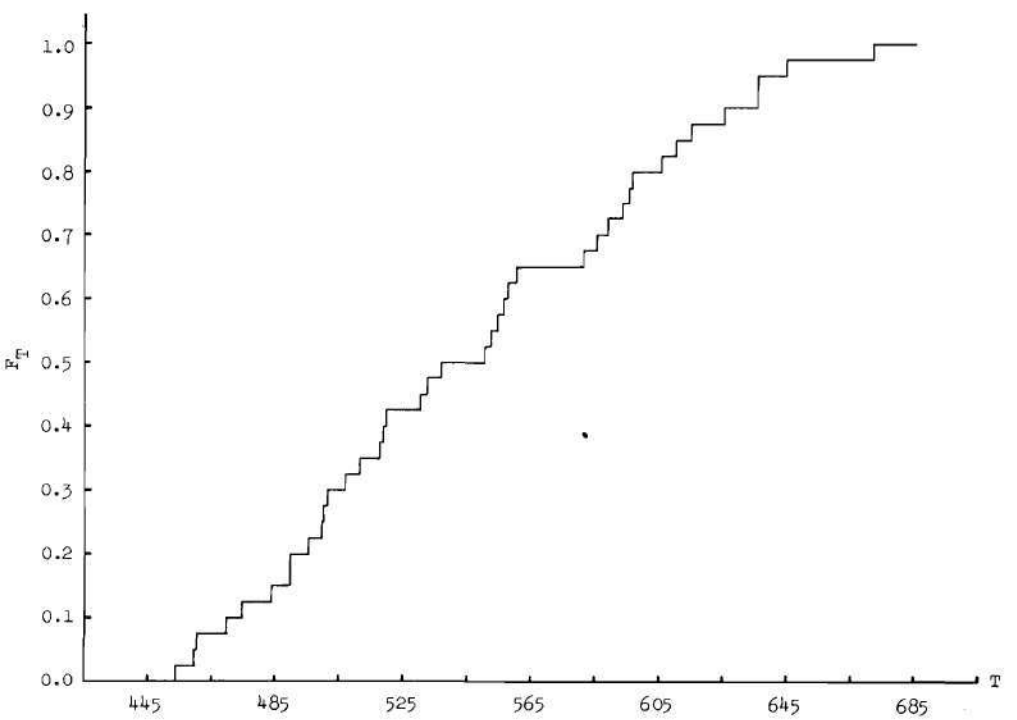


Figure 40. Distribution Function for T in Milliseconds, Master Set.

from a consideration of the histogram, can be fairly well approximated by a straight line.

The numeric and decibel values of word power are distributed as shown by F_p and F_P of Figures 41 and 42, respectively. Again, the shapes of these curves correlate well with those of the corresponding histograms. The shape of the F_p curve is similar to one previously obtained (6) for the fifty monosyllables of PB 1.

Distributions were not plotted for the values of noise power over each word-length segment of noise, i.e., for P_n, but consideration of tabulated values reveals that spread in values of noise power is quite small, being only 0.58 db. Further consideration of values measured for "live" (unrecorded) noise indicates even less variation over word-length intervals than for the recorded noise actually used. Much of the variation in P_n can be attributed to magnetic tape coating irregularities, indicating that noise power, in the "live noise" case, can be considered as being practically constant and contributing very little to the spread in word signal noise ratios. For speech items of shorter duration than the monosyllables used here, or for types of noise different from the band-limited white gaussian type used, the noise power cannot be considered constant.

When values of word energy w and duration T are arranged in increasing order, a weak correlation between these two parameters is evident. That is, to some extent words having shorter durations tend to have smaller energy, although many exceptions exist. This correlation seems to be most pronounced for low-energy words, as can be seen from Table 10, where the first 10 words in each ranking are shown, and where energy and

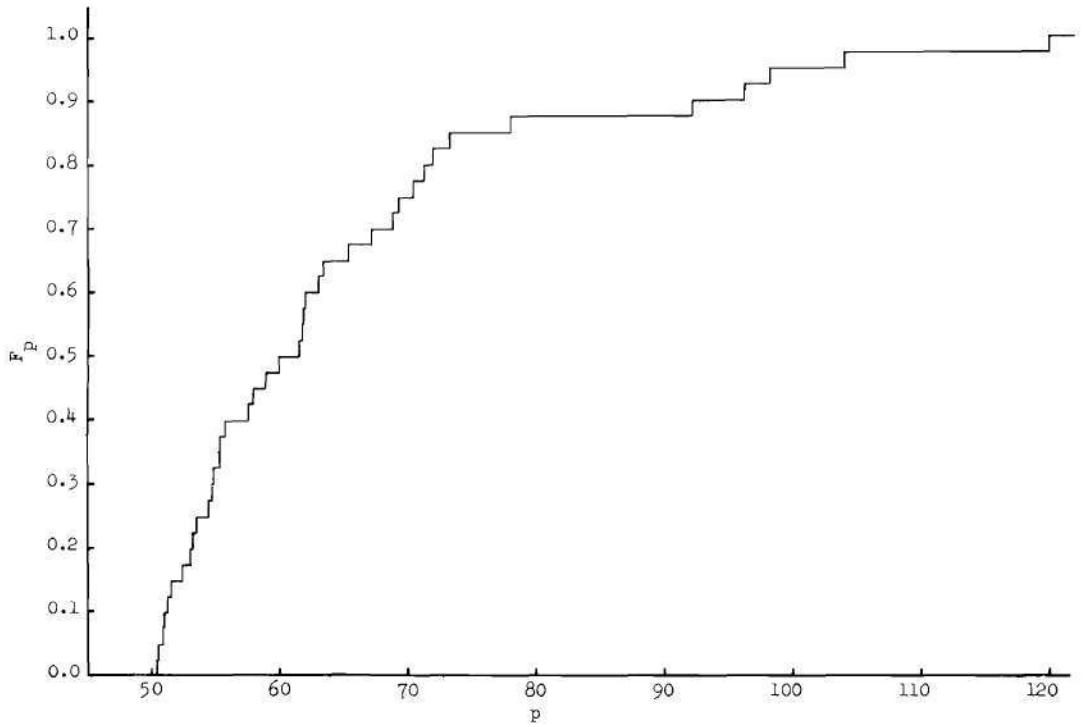


Figure 41. Distribution Function for Word Power p, Master Set.

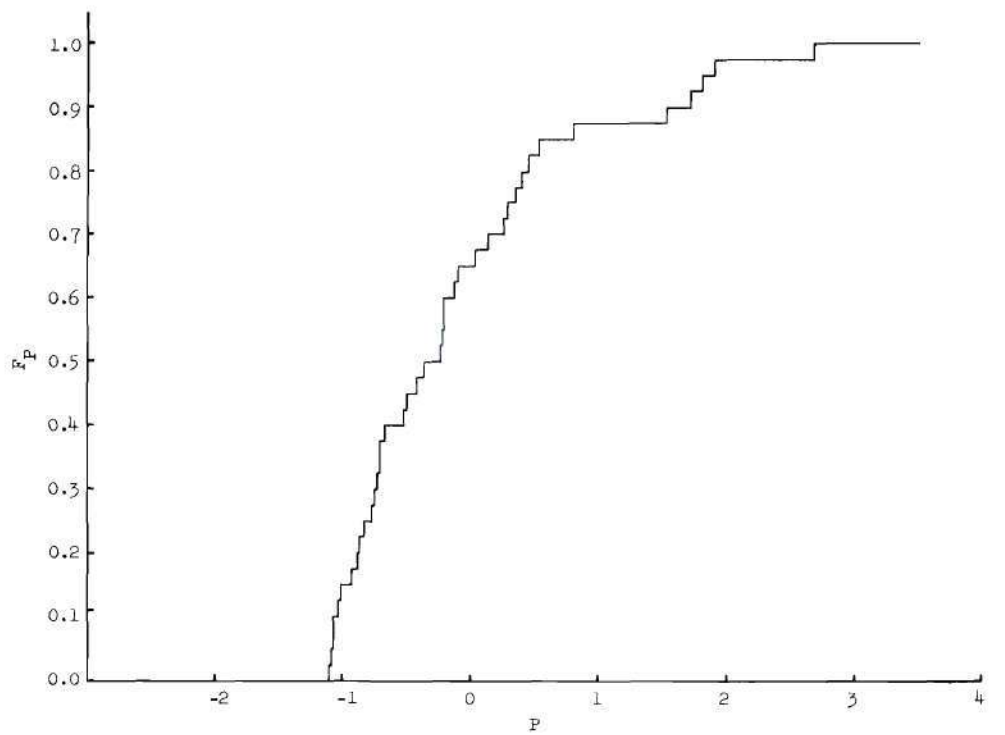


Figure 42. Distribution Function for P in DB, Master Set.

duration increase with rank number. As can be seen, eight of the first ten words are common to both rankings. Further examination of the data indicates that only part of the variation in energy is attributable to variation in duration. Other factors, such as the inherent relative intensity of the different vowel sounds (34), undoubtedly contribute to the variation.

Table 10. Ranking of Words by Energy and Duration

By Energy		By Duration	
rank	word	rank	word
1.	fig	1.	ache
2.	rate	2.	neck
3.	neck	3.	rate
4.	ache	4.	fig
5.	check	5.	muck
6.	take	6.	path
7.	muck	7.	take
8.	who	8.	check
9.	turf	9.	turf
10.	lush	10.	deck

Reliability of Scores

Although no detailed investigation was made of the statistical accuracy of articulation curve points, the 90 per cent confidence interval was calculated in several cases. As already pointed out, the reliability of the mean score (averaged over several tests or several listeners

or both) increases with the number of tests and/or listeners, and is greatest near the low and high ends of the articulation curve. Near the 50 per cent level, where scores have the least reliability, the data was examined for several arbitrarily-chosen listener/word-set combinations, and 90 per cent confidence intervals calculated. The results including the confidence interval expressed as a symmetric percentage interval about the mean score, are given in Table 11.

Table 11. The 90 Per Cent Confidence Intervals for Several Articulation Curve Points

Line	Listener	Word Set	Mean Score in Per Cent	Confidence Interval	Confidence Interval as a Percentage of Mean Score
1.	Team	Master Set	55.7	55.7±3.18	55.7 ± 5.7%
2.	Team	Subset G	50.4	50.4±5.85	50.4 ± 11.6%
3.	Team	Subset H	50.1	50.1±5.84	50.1 ± 11.7%
4.	NS	Master Set	51.5	51.5±5.70	51.5 ± 11.1%
5.	NS	Subset H	48.0	48.0±10.94	48.0 ± 22.8%

From values in Table 11, one can make statements such as, "With 0.9 probability, the "true" mean score lies within ±5.7 per cent of the observed mean score (value used in plotting curve), for the point nearest the 50 per cent level on the team articulation curve for the master word-set." These "worst case" values set a rough upper limit on percentage possible error (with probability 0.9) when using the observed mean score for plotting a point on the curve. For example, in the case just cited, the

maximum possible such error is about ±6 per cent, that is, with 90 per cent confidence one would expect the plotted point to be no more than 5.4 per cent below, or 6.05 per cent above, the true value.

From lines 1, 2, and 3 of Table 11, it is clear that the uncertainty in the true value is roughly twice as great for subset team curves as for the master-set team curve, this being due to having averaged over more repetitions in the latter case. Averaging over only one listener instead of three also increases the uncertainty (lines 1 and 4), as does decreasing the number of repetitions for a single listener (lines 4 and 5). The raw scores, from which mean scores are computed, also possess a varying degree of uncertainty, depending upon whether 40 words or 20 words are involved. Based on these few calculations, the scores seem to be satisfactorily reliable. In the worst case tabulated, one could expect, with 90 per cent confidence, that the "true" value of mean score lay between roughly 38 and 58 per cent, whereas the value used for plotting was 48 per cent.

Distribution of Subjective Word Parameters

In addition to being useful in the prediction schemes, the subjective word parameters of threshold and spread are of interest in themselves, at least to the extent that they describe the intelligibility properties of the words. When viewed in this way, only "valid" values of these parameters should be considered, i.e., only those α's and Δ's which represent good estimates of the NMGF threshold and spread, as discussed in Chapter III. A fairly wide range of values were encountered for α and Δ, including variations from listener to listener (for the same word) and from word to word (for the same listener). For listeners

JB, WN, and NS, respectively, the ranges covered by valid values of α are 14.9, 18.4, and 12.9 db. Thus, to NS, the words seemed to be grouped more closely in intelligibility than to WN, and hence possessed a greater "density" over their smaller range along the SN axis. This fact is reflected in the larger slopes observed for the actual and predicted master-set articulation curves for NS (see Table 9). The ranges between highest and lowest values of Δ were 15.3, 12.6, and 16.7 db for JB, WN, and NS, respectively. Thus the words seemed more "alike" in spread to WN than to the other listeners. The smallest Δ observed was the same for each listener, namely, 4 db, the above differences being due entirely to different maximum values for the three listeners. On this basis, one would expect the master-set curves for WN to be more "spread out," i.e., to have smaller slopes. That this is true can be seen from Table 9.

Considering the combined set of valid thresholds for all listeners, the master-set words have a threshold distribution approximated by the histogram of Figure 43. In constructing this histogram, 7 of the 120 values of α were discarded as being "non-valid." The mean value of all the thresholds is indicated by a dashed line. Again considering all valid values (109 of 120) of Δ for the 40 words and three listeners, the distribution of spreads was determined and plotted as the histogram of Figure 44. From a comparison of Figures 43 and 44, it is evident that a strong central tendency exists in the α-histogram, showing that values of threshold tend to cluster around the mean value. The symmetry about the mean value is less pronounced in the Δ-histogram, but the apparent skewness may be due to the small amount of data involved or to the quantization

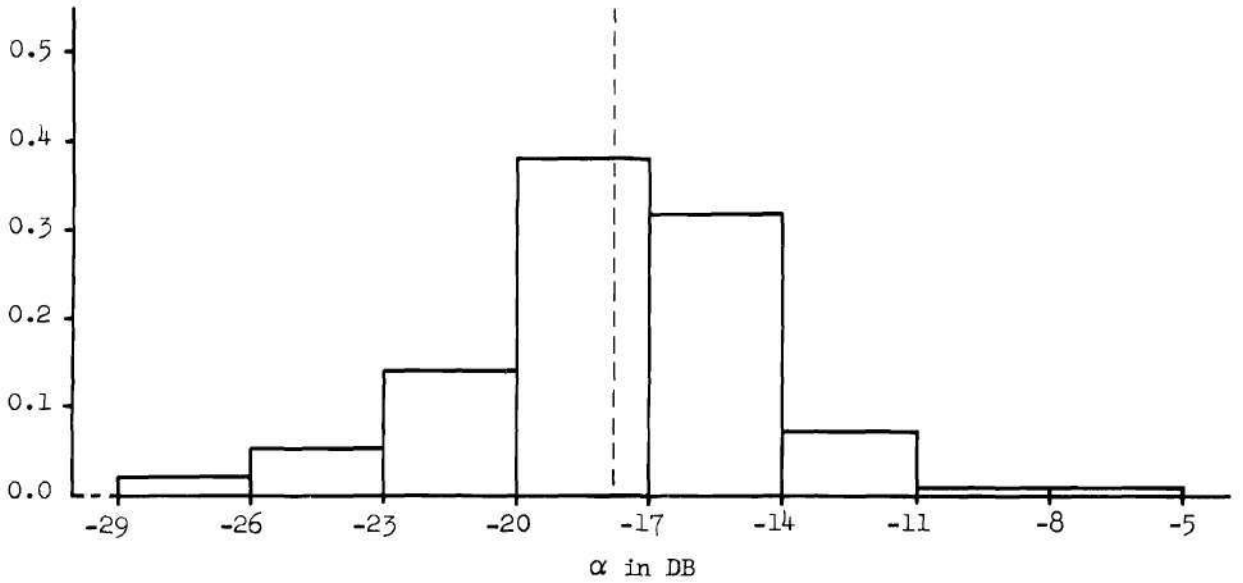


Figure 43. Histogram of Valid Values of Threshold, for All Listeners.

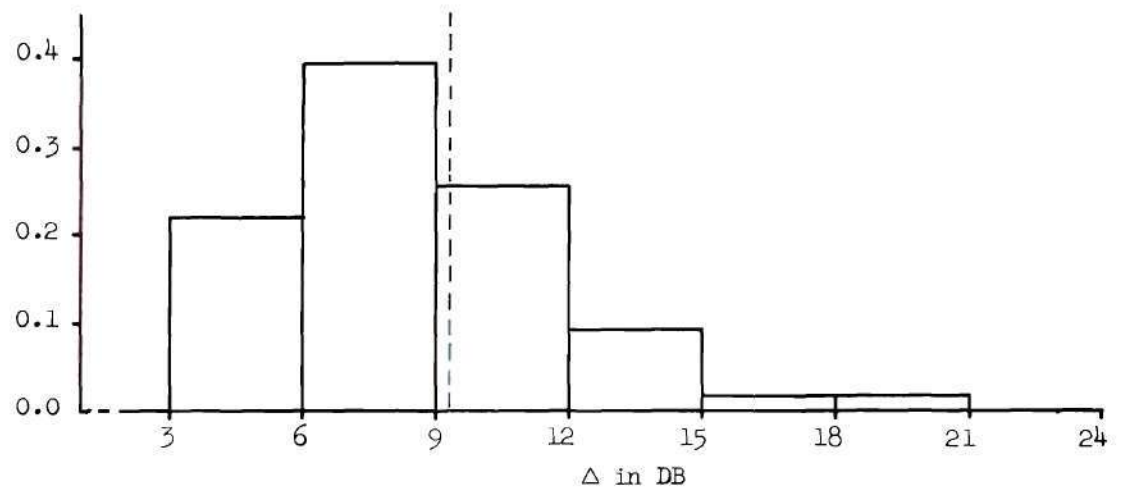


Figure 44. Histogram of Valid Values of Spread, for All Listeners.

of Δ's resulting from the method of selecting points for least squares fitting of the NMGF's. It is apparent, from these figures, that neither thresholds nor spreads are uniformly distributed, and that the intelligibility characteristics of words are quite varied, even for a relatively small set of words.

When words are ordered by threshold or spread for the three listeners, no marked or consistent correlation can be observed between a word's phonetic structure and its rank order in α or Δ. Although one would expect a relationship between phonetic structure and intelligibility, this relationship in the present case is apparently masked by the effects of context, i.e., the fact that the intelligibility of a word is strongly dependent upon the choice of other words in the test set. For example, there are six words in the master set possessing "æ" (as in "path") for the vowel sound, and only two words possessing "ɪ" (as in "fig") for the vowel sound. Assuming that the vowel sound is the first to be recognized as SN is increased, a listener clearly must choose between six or two alternatives when the sound is æ or ɪ, respectively, implying a higher probability of making the correct choice in the latter case. Hence one would expect the words containing ɪ to be more often guessed correctly, and hence to give higher scores at a given SN. This was substantiated by computing the average threshold $\bar{\alpha}$ over all listeners and all words containing æ, resulting in an $\bar{\alpha}$ of -15.8 db, and repeating the process for ɪ, resulting in an $\bar{\alpha}$ of -19.7 db. Thus the six-word subgroup was some 4 db less intelligible, in terms of threshold, than the two-word group. Considering n-word subgroups, where n ranged from six (æ) to one (ɑ, ɪu and ɔɪ), and averaging over all subgroups having

the same number of words n, the average threshold was found to vary with n as shown in Table 12.

Table 12. Variation of $\bar{\alpha}$ with Size n of Common Vowel-Sound Groups

n	6	5	4	3	2	1
$\bar{\alpha}$	-15.8 db	-17.9 db	-18.3 db	-17.0 db	-20.0 db	-18.9 db

A general trend to lower thresholds with decreasing n is evident from Table 12, although the variation is undoubtedly affected by other considerations, such as the variation of inherent intelligibility among the vowel sounds themselves. The trend is most clear-cut in the cases where the number of words available for averaging was greatest, i.e., for n = 6, 5, and 4. Only three words were averaged for n = 1. One reasonable conclusion which can be drawn from these results is that context had a marked effect on the relative intelligibility of various words in the present case, and that this effect is difficult to separate from the effect of phonetic structure.

While considering the α and Δ parameters, the question naturally arises as to whether there is any relationship between them, i.e., whether a low value of α implies a large Δ, or whether the reverse is true. Some idea of the interdependence of these parameters can be obtained by ordering words by threshold, tabulating the corresponding values of Δ, and averaging the values over two-db increments of α. When this is done and the results plotted for each listener, the results appear as in Figure 45. From these curves, a weak trend in the direction of smaller Δ's

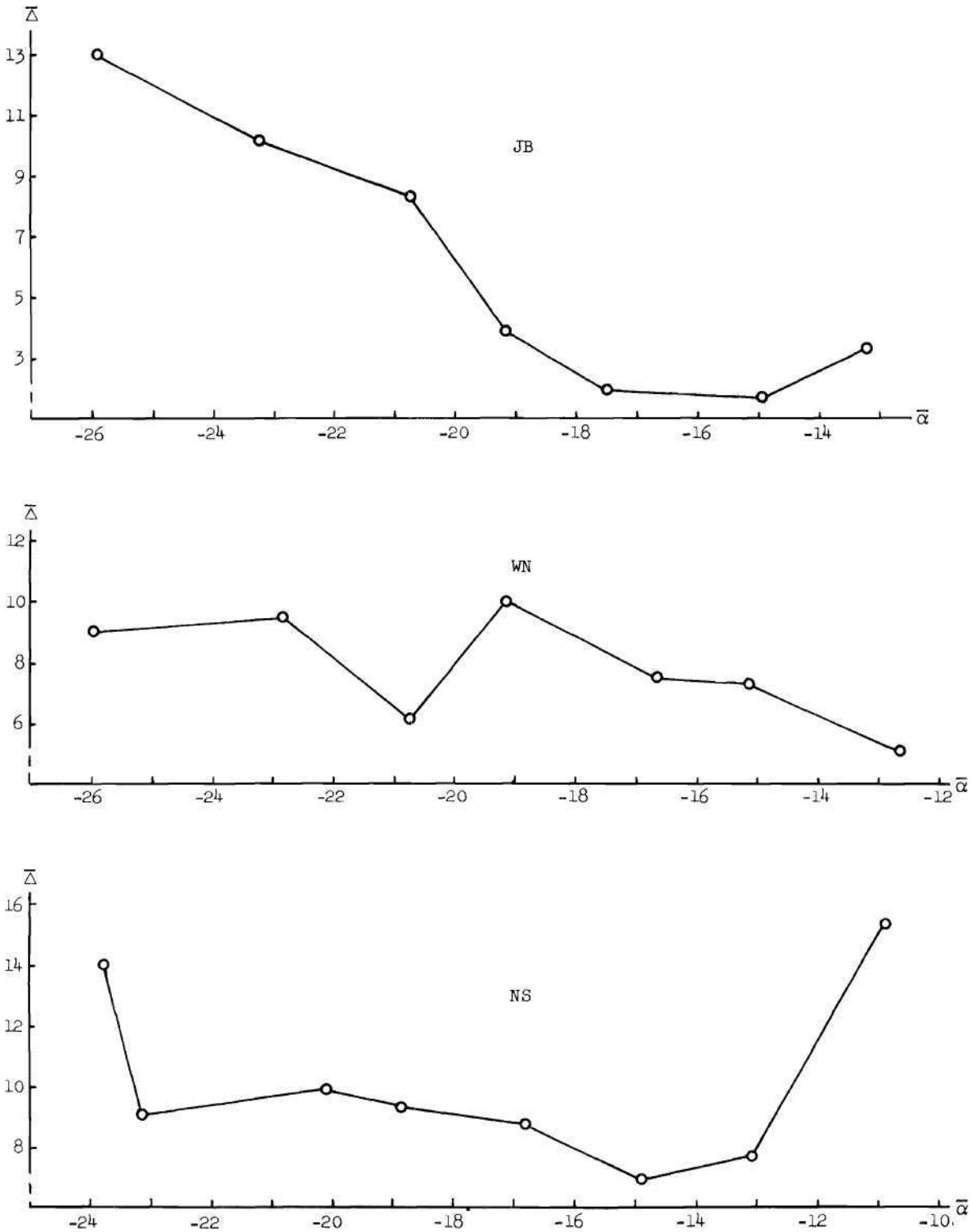


Figure 45. Average Δ versus Average α for JB, WN, and NS.

for increasing α's is evident, this trend being most pronounced for listener JB. This implies that high threshold (less intelligible) words tend to have smaller NMGF spreads or, what is the same thing, larger NMGF slopes. This trend is also evident in Figure 46, in which results have been averaged over all listeners, except for the right-most point. This point is due entirely to the one word "class" as heard by NS, this word being the only one in the final averaging interval. For this reason, not much significance can be attached to this point or to the sharp reversal of trend associated with it. If this point is ignored, the curve implies a weak trend toward smaller spreads with increasing threshold. That is, the curve implies an increase in NMGF slope with increasing threshold. This result is contrary to that obtained for alphabet letters by Curry, Fay, and Hutton (22), who state that there is "...a strong inverse relation between slopes and 50 per cent intelligibility levels."

In the prediction scheme based on step-approximations to the NMGF's, the use of the β parameter is quite tedious when team curves are to be plotted for a large group of listeners. The procedure involves predicting an articulation curve for each listener and then averaging the curves. A great deal of labor could be saved if thresholds could be averaged over all the listeners and then the prediction scheme applied only once to yield the team curve. The question involved here is whether the thresholding and averaging operations are commutative, and the answer is clearly in the negative because of the non-linear nature of the thresholding operation. This can be shown easily by postulating a one-word test with two listeners: the average of the step-predictions would be a two-step curve (assuming the listeners had different thresholds for the word),

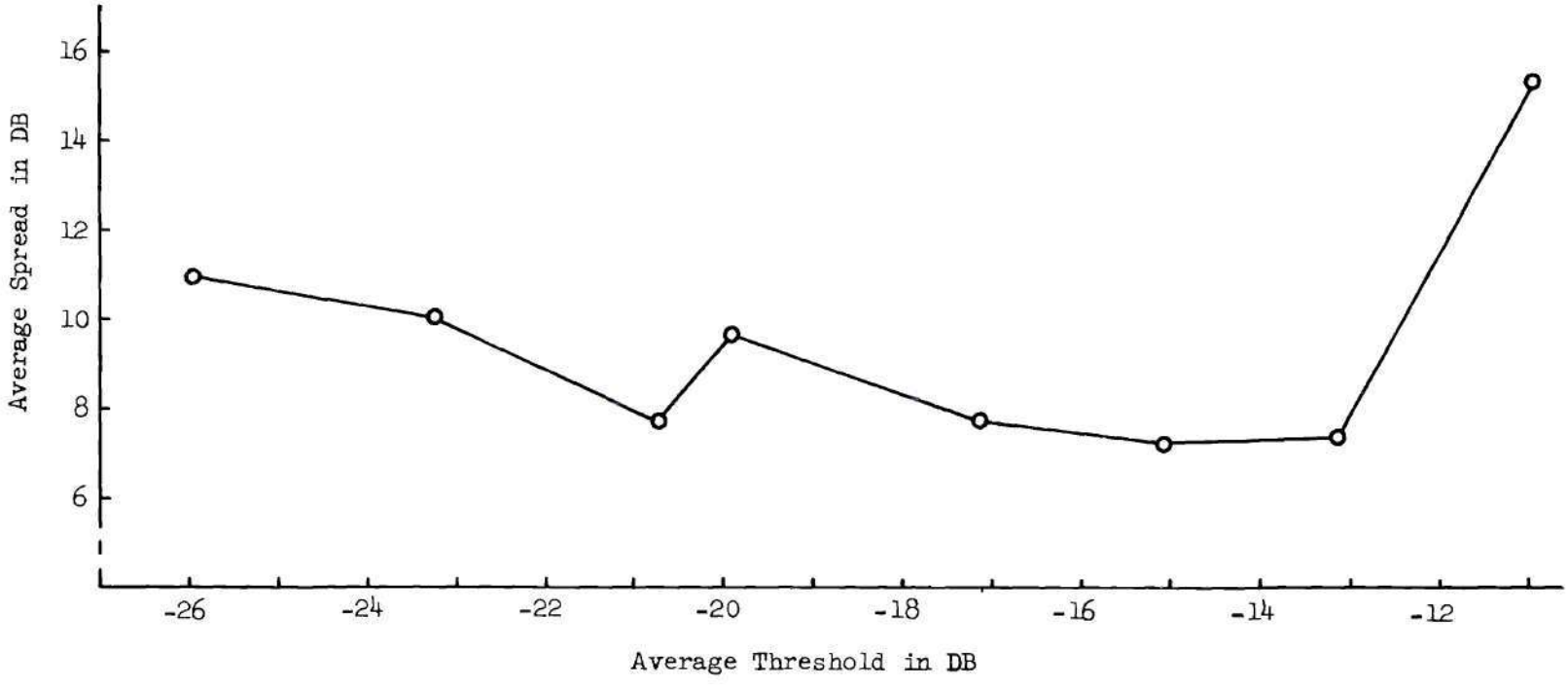


Figure 46. Average Spread versus Average Threshold, for Team.

whereas a step-prediction based on the average threshold would be a one-step curve. In the present case it is clear that the use of 40 values of average threshold will yield a more granular curve (40 steps) than the average of the three 40-step individual curves (120 steps). Nevertheless, for a sufficiently large number of listeners the use of average thresholds may result in essentially the same predicted curve at a considerable saving in effort. To evaluate this possibility, a team curve was predicted by the step-approximation procedure, using values of β which were averages over the listeners. This curve, as discussed in Chapter II, is essentially a plot of $F_{\overline{\beta}}$, where $\overline{\beta}^i$ is the average threshold for the i^{th} word. For comparison, points on the previously calculated team step-prediction (see Figure 21) were plotted on the same graph; this curve is essentially $\overline{F}_\beta$, the average of F_β for the three listeners. The results, shown in Figure 47, show good agreement between $F_{\overline{\beta}}$ and $\overline{F}_\beta$. A single curve has been fitted to the combined sets of points to emphasize the close agreement. Apparently, for as few as three listeners and 40 words, the use of average thresholds in the step-prediction of team curves is permissible.

Comparison of Listeners

From a comparison of the various articulation curves and NMGF's, it is evident that considerable differences exist between the listeners in their responses to particular words. The simplest way of comparing one listener to another is to consider the values of threshold and spread associated with the two listeners for a given word. For example, there is a wide variation of threshold among the listeners for the word "fig," the values of α being -26.1, -18.3, and -15.3 db for JB, WN, and NS respectively. Likewise, JB, WN, and NS exhibited values of Δ, for the word

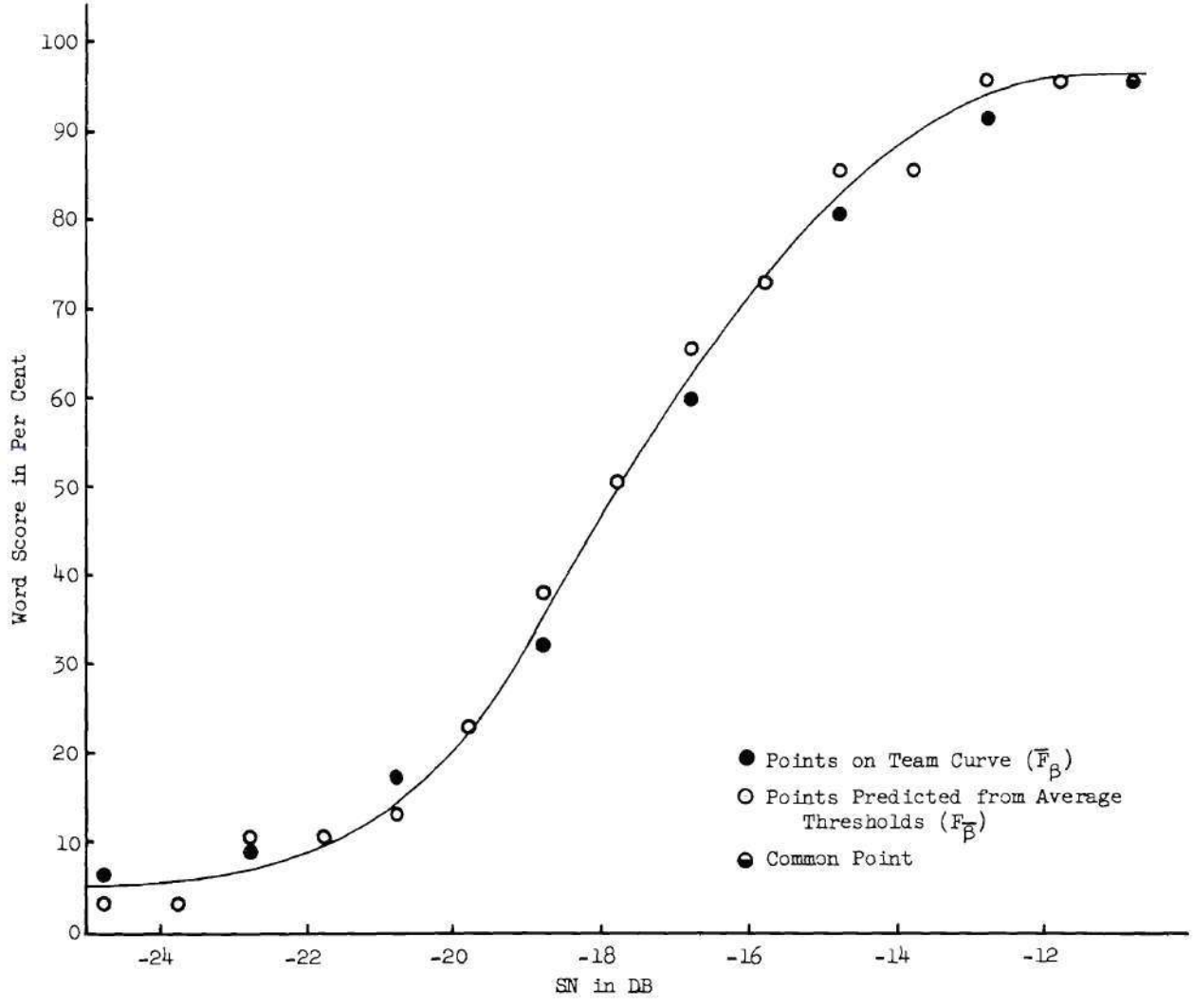
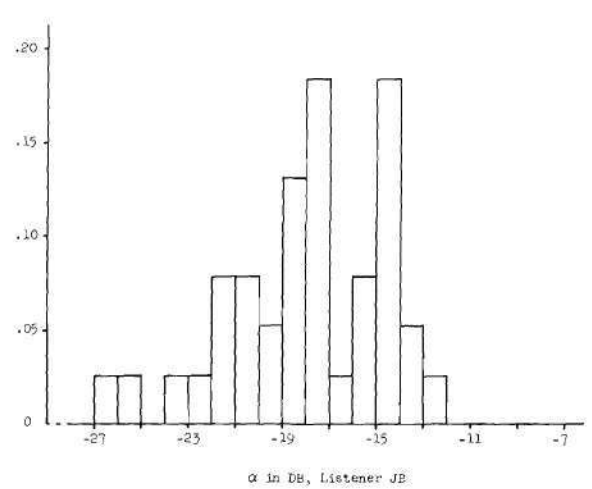


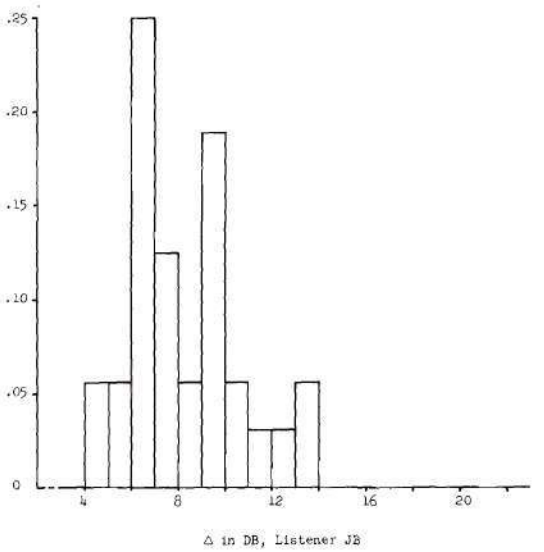
Figure 47. Comparison of Average Prediction and Prediction Using Average Thresholds.

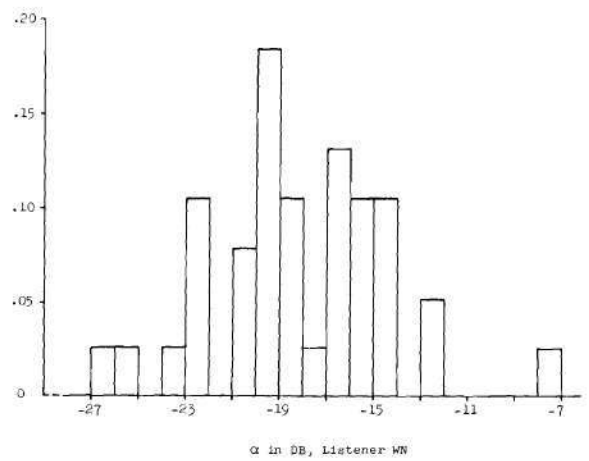
"class," of 7.7 db, 12.1 db, and 15.4 db, respectively. Some of this variation is accounted for by differences in hearing acuity, "listening ability," intelligence, and motivation. Even if such factors could be equalized for all listeners, there would undoubtedly remain some variations in response, in the sense that one listener might find particular words difficult while another might find the same words relatively easy.

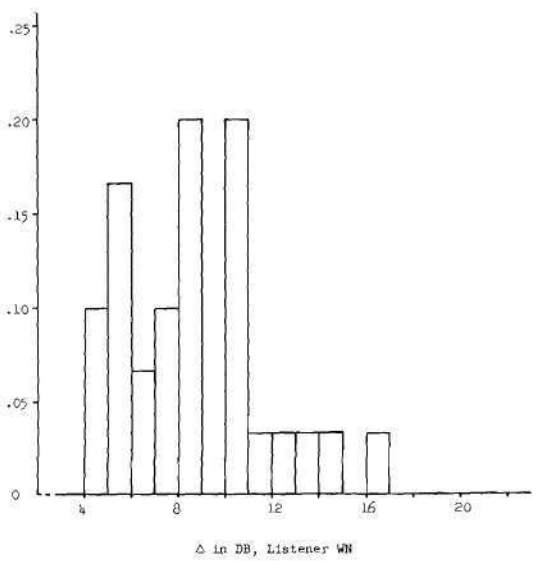
A gross comparison of listener performance can be made on a basis of the average threshold, or average spread, for all words. A more detailed comparison can be made by considering the distribution of thresholds and or spreads for the three listeners, as displayed by the six histograms of Figure 48. The histograms for α indicate that NS, in general, had slightly higher thresholds than did JB or WN, while WN had the largest range of thresholds. Also, from the shape of the curves, the thresholds for NS seem to be clustered more densely in the central region of the curve, which would explain the larger slope, for the step-predicted master set curve, observed for NS. The average threshold, considering all valid values of α, was found to be -18.0 db for JB, -18.2 db for WN, and -17.2 db for NS. The Δ-histograms indicate that JB had a "tighter" distribution of Δ's than either WN or NS, while NS displayed the largest range of Δ's. The average Δ's for JB, WN, and NS are, respectively, 8.2 db, 8.3 db, and 8.8 db, reflecting the slight preponderance of large spreads for NS.

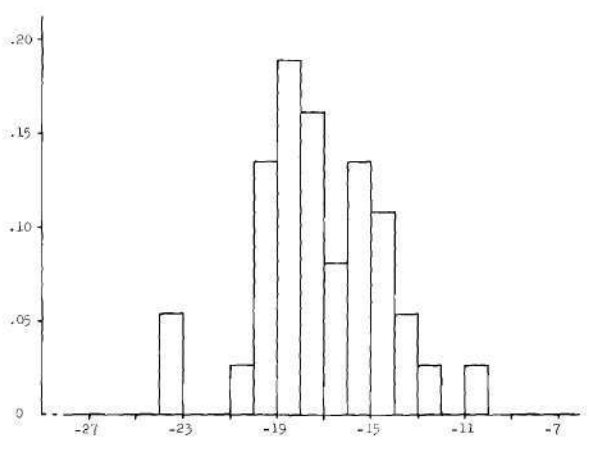
At least on a basis of average threshold and spread, the listeners are not too dissimilar, the small differences being in the detailed structure of their α and Δ distributions. This is displayed most accurately by the cumulative distribution functions of Figures 49 and 50, which were











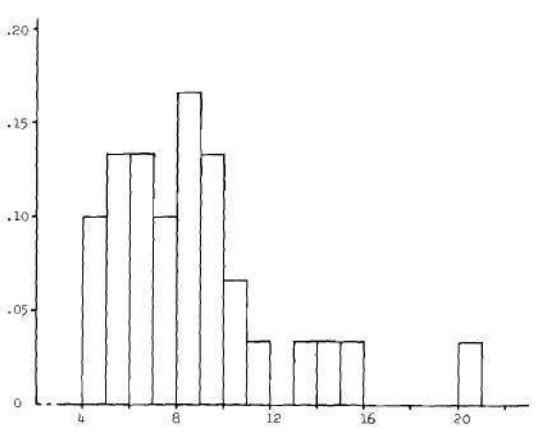


Figure 48. Histograms for Threshold and Spread, Various Listeners.

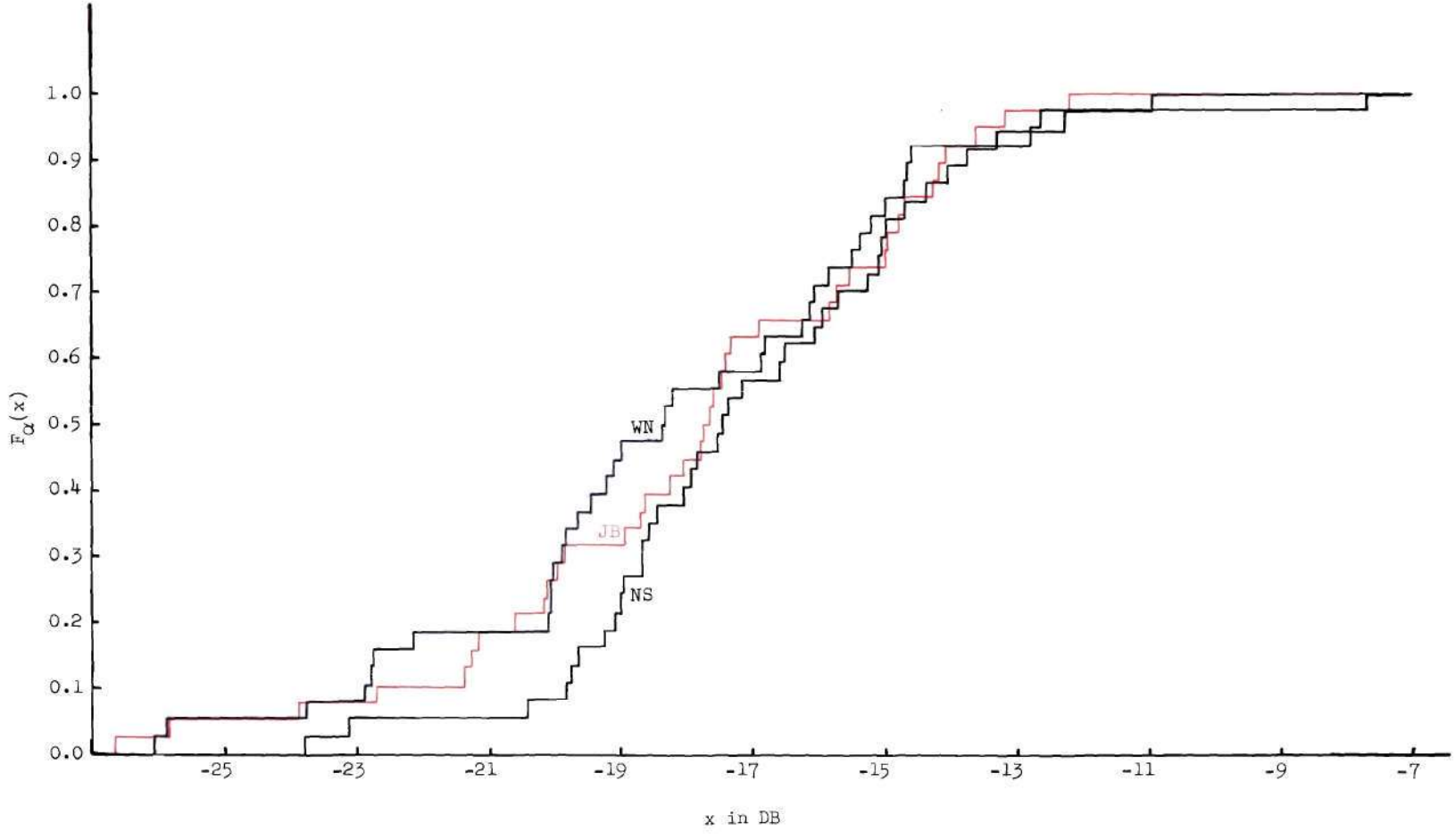


Figure 49. Threshold Distribution Functions for JB, WN, and NS.

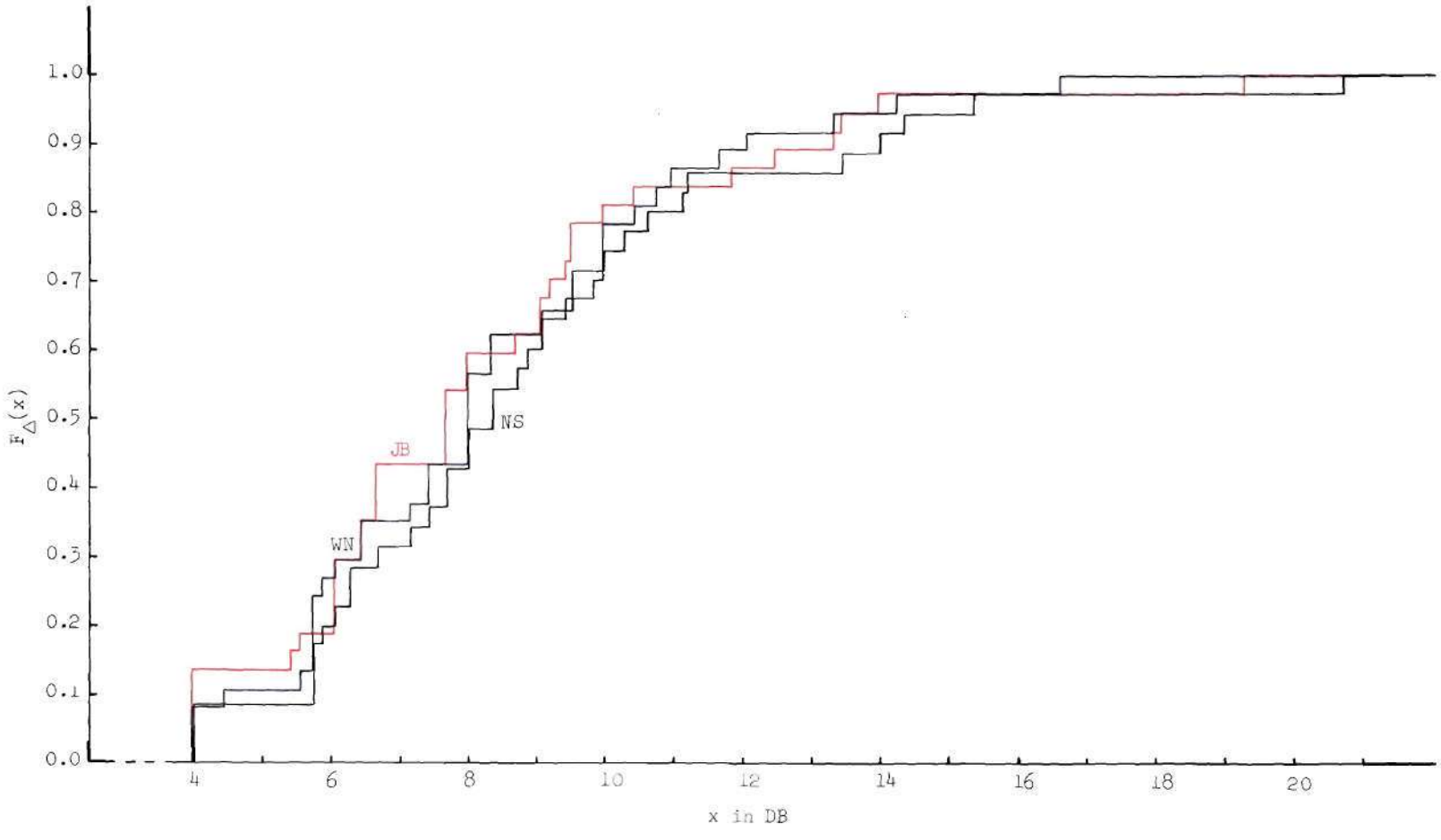


Figure 50. Spread Distribution Functions for JB, WN, and NS.

plotted for all valid values of α and Δ, respectively. In each of these figures, the red curve is for listener JB, the dark-blue curve is for WN, and the black curve is for NS. These functions, denoted by F_α and F_Δ, are seen to be quite similar to one another in shape and location, the major differences being found at small values of threshold and at large values of threshold and spread. This indicates that the general responses of the three listeners to the master set, without taking into account particular words, were quite similar.

From the foregoing, one can conclude that if there are any major differences between listeners, they must lie in the variation of response to particular words, and not in the general response to the word-set. It would be possible, for example, for the listeners to have identical F_α's but to have completely different orderings of this parameter, i.e., words which had low thresholds for one listener might have high thresholds for another, and vice versa. Clearly, three possible listener-pairs must be considered in making any such comparisons, and only words which have valid parameters for all three listeners can be used. A rough comparison of the ordering sequences associated with the α's and Δ's of the listeners can be made by taking the differences in rank-order and plotting from these a histogram. If the ranking by threshold, for example, produced the same sequence of words for listeners number 1 and 2, then the two order-numbers associated with each word would be identical, and all differences in these paired order-numbers would be zero. If the difference is denoted by α_{12}, then the histogram of α_{12} would consist of a single bar enclosing the value "zero," indicating excellent correlation between the ranking of thresholds for listener number 1 and for listener

number 2. One could conclude, from such a histogram, that if a word has a low threshold for listener 1, it also has a low threshold for listener 2, and conversely for high threshold words. In fact, in this case of complete identity in response, one could say that the word having the fifth lowest threshold for number 1 also has the fifth lowest threshold for number 2, and, in general, the word of rank "i" for listener 1 also has rank "i" for listener 2. Words having a low rank for listener 1 and a high rank for listener 2 (or vice versa) would exhibit large α_{12}'s, and these large values would cause the histogram to spread out. In the "worst" case, where the ordering sequences were exactly reversed, and if differences are taken in a consistent direction (thus yielding both positive and negative values of α_{12}), then the histogram would be "flat" or uniform in height, and centered at zero. This would indicate that low-threshold words for one listener were high-threshold words for the other, and vice versa, while median-ranked words for one listener were also median-ranked for the other. The actual case is somewhere between these two extremes, as explained below.

With listeners JB, WN, and NS being identified by the numbers 1, 2, and 3, respectively, the three rankings of words by threshold were examined, word by word, to obtain the differences α_{12}, α_{23}, and α_{31}, where

$$\alpha_{jk} = (\text{rank order for listener k})-(\text{rank order for listener j}). \quad (69)$$

A complete description is obtained by considering only the j,k pairs 1,2; 1,3; and 3,1; since α_{jk} is simply the negative of α_{kj} and since the α_{jj} and α_{kk} are trivial. Histograms of these three differences are

shown in Figure 51, where the α_{12} histogram compares JB and WN, the α_{23} histogram compares WN and NS, and the α_{31} histogram compares NS and JB.

From Figure 51, it can be seen that a strong central tendency exists for α_{23} and α_{31}, indicating that most words are close in rank-order for WN and NS, and also for NS and JB. This tendency is less marked for α_{12}, indicating that the words differ more in order of intelligibility for JB and WN. The two listeners who respond most nearly alike are WN and NS; for this pair of listeners over 36 per cent of the words differ in rank-order by 2 or less, compared to the largest possible difference (in rank-order) of 35 for the 36 words used. For NS and JB, over 30 per cent of the words differ by 2 or less in rank-order, while for JB and WN less than 17 per cent of the words differ by this amount. One can conclude that, for WN and NS, the "intelligibility" is distributed over the various words in a very similar manner, and that this is true, to a lesser extent, for NS and JB. At the same time, the intelligibility rank of a given word for JB does not imply a great deal about the rank of that word for WN. Note that in these comparisons the absolute values of threshold have no bearing, the "closeness" of two listeners' responses to a given word implying only that the relative position of the word in intelligibility is similar.

In a similar manner, the histograms for Δ_{12}, Δ_{23}, and Δ_{31}, where Δ_{jk} is defined in a manner analogous to α_{jk}, were plotted as shown in Figure 52. In this case, the listeners most alike in terms of NMGF spread (or slope) were JB and WN, while the most dissimilar listeners were NS and JB.

Finally, the appropriateness of the linear approximations to the

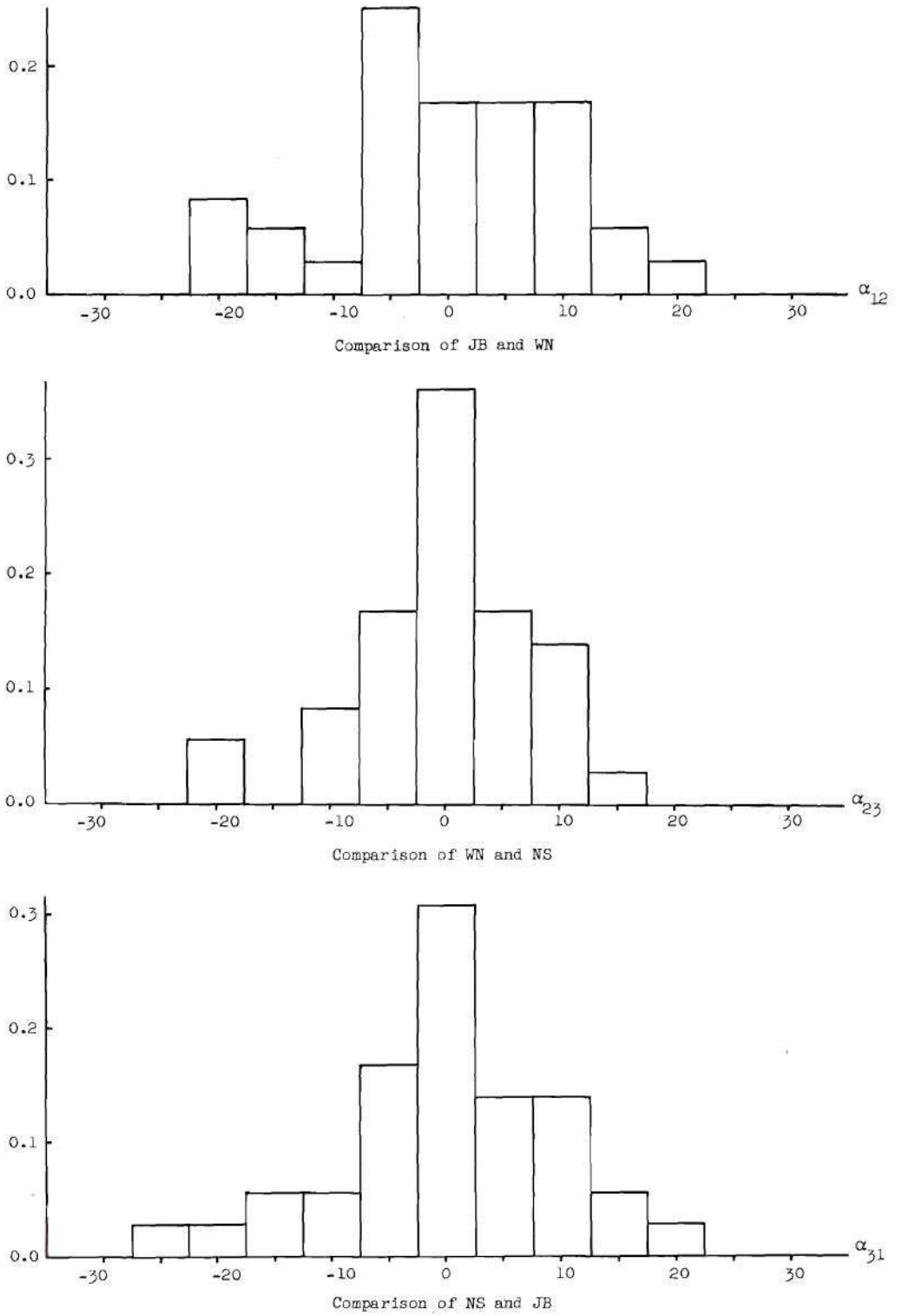
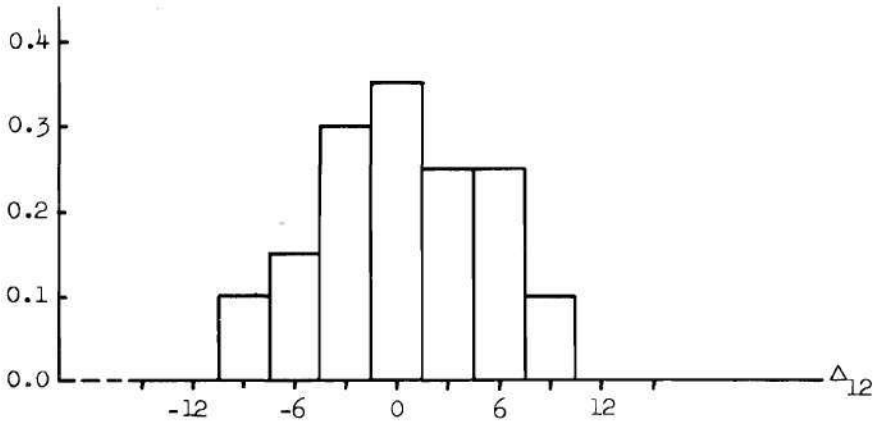
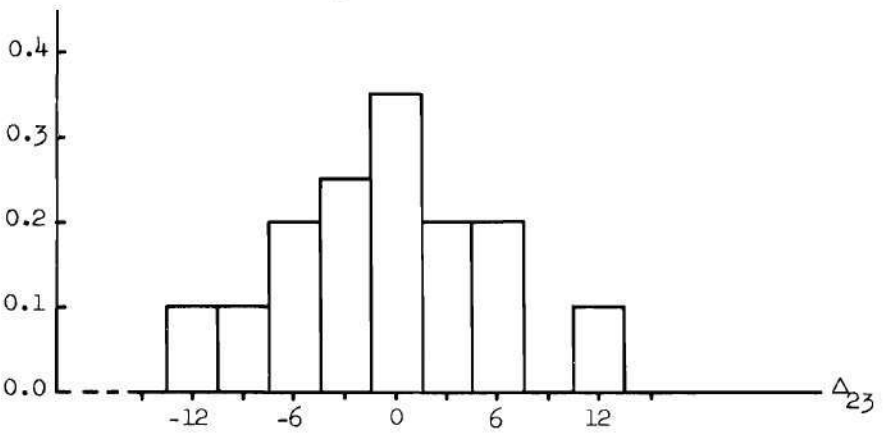


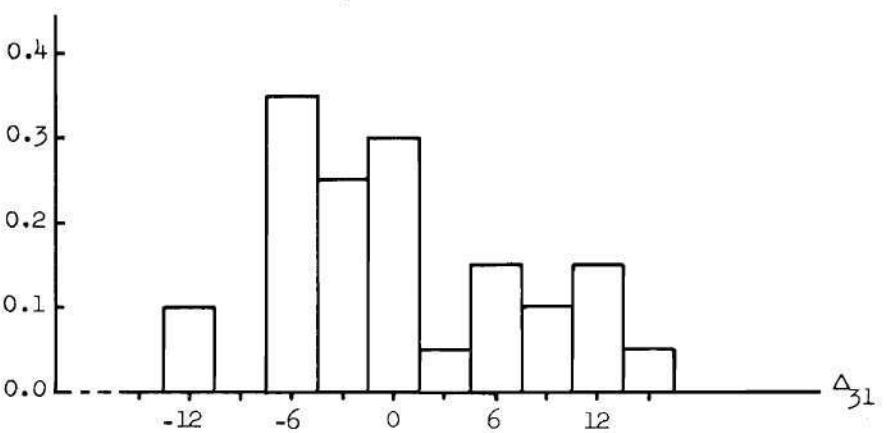
Figure 51. Histograms for Rank-Order Differences in Threshold.



Comparison of JB and WN



Comparison of WN and NS



Comparison of NS and JB

Figure 52. Histograms for Rank-Order Differences in Spreads.

NMGF's was investigated for each of the three listeners. This was done by plotting histograms for the magnitudes of the linear correlation coefficient $|r|$ associated with the sets of NMGF points. The histogram for each listener is shown in Figure 53, from which it can be seen that WN has the largest percentage of high $|r|$ values. The percentage of $|r|$'s equal to or greater than 0.95 is 42.5, 60, and 57 per cent, for JB, WN, and NS respectively. To the extent that high values of $|r|$ indicate that straight lines are good approximations to the NMGF's, these histograms show that the linear approximations were somewhat better representations for WN and NS than for JB. This is confirmed by the linear prediction of the master-set articulation curve (see Figures 18, 19, and 20), this prediction being closer to the actual curve in the case of listeners WN and NS than in the case of listener JB.

It is clear that the accuracy with which the NMGF's are approximated by straight line segments, as roughly measured by $|r|$, affects the validity of both prediction schemes. It is also reasonable to expect words having small values of spread Δ to give better results than large-Δ words in the step-prediction scheme, since this scheme assumes $\Delta = 0$. The quantity $\frac{|r|}{\Delta}$, then, can reasonably serve as a criterion of suitability for the use of a given word in the step-prediction scheme, large values of this quantity indicating more suitability than small ones. The values of this quantity were calculated for all 40 words and for each listener, with the result that the average value of $\frac{|r|}{\Delta}$ was 0.129, 0.121, and 0.114, for JB, WN, and NS, respectively. One might thus expect the step-prediction scheme to be more accurate for JB and WN than for NS, a fact which is confirmed by comparing actual and step-prediction curves in Figures

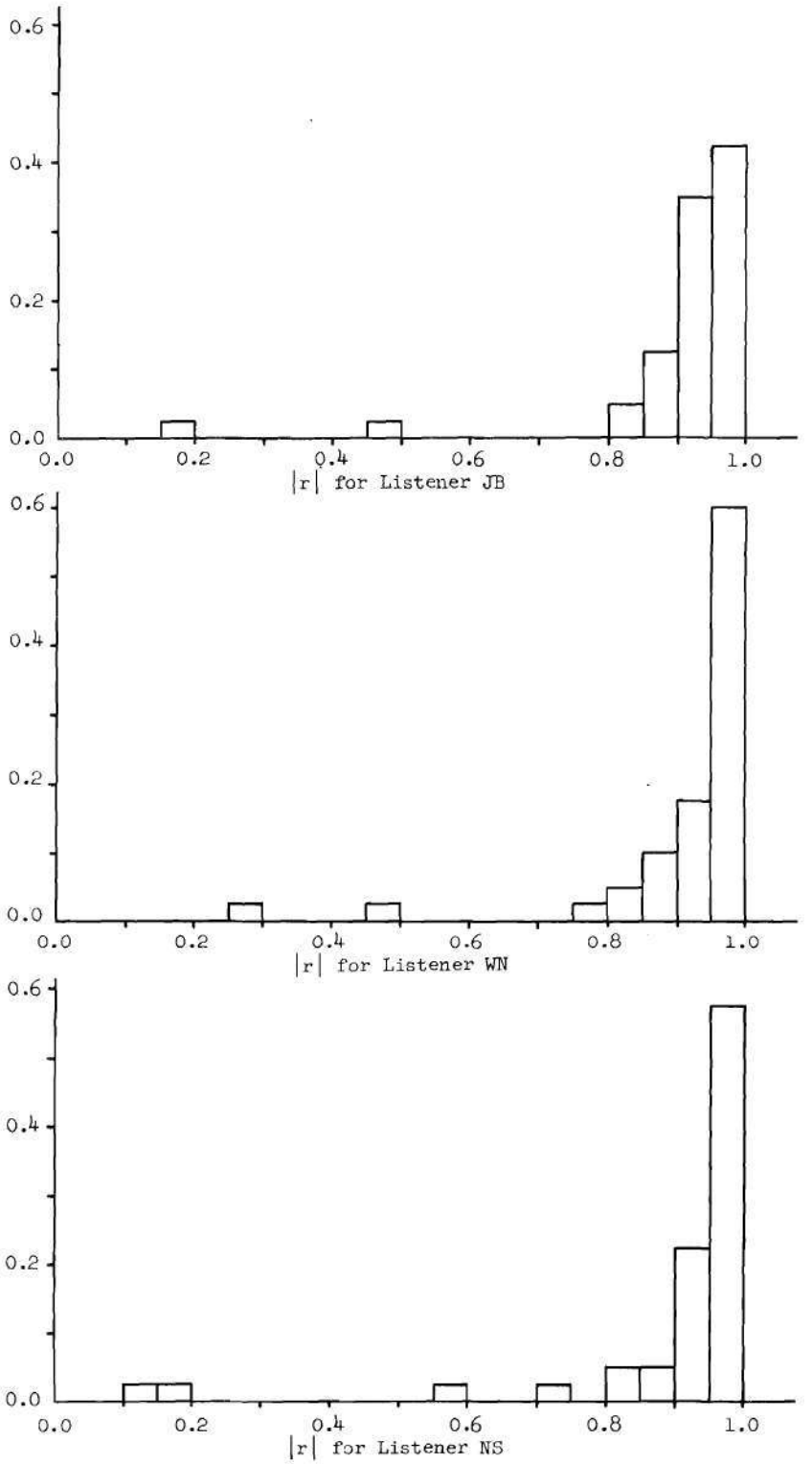


Figure 53. Histograms for Magnitude of Correlation Coefficient.

18, 19, and 20. As can be seen from the curves, the listeners rank in prediction accuracy in the same order that they rank in average $\frac{|r|}{\Delta}$.

CHAPTER VI

CONCLUSIONS

From the results of subjective and non-subjective measurements on a set of noise-masked mono-syllabic words, it has been shown that such words may be characterized by three basic parameters. One of these, the word power, is a physical attribute of the speech waveform; while the remaining two, threshold and spread, are attributes of the subjectively-measured noise-masked gain function. The existence and measurability of these parameters has been demonstrated, and their usefulness has been illustrated for a particular set of test conditions. Certain conclusions, valid for the given test conditions but intuitively applicable for larger numbers of listeners or test words, can be drawn from the results, as follows:

1. The results of conventional articulation tests can be presented in a novel way, i.e., the set of thresholds and gain function spreads, which, together with values of noise power and individual word power, contains essentially all of the information in the standard articulation curve. It follows that these parameters play a basic role in determining the intelligibility of a set of noise-masked test words.

2. For a given set of words and masking noise, the shape and location of the articulation curve depend on, and can be predicted from, the basic word parameters. The major determinant of the articulation curve is, for a given noise, the set of ratios of threshold to normalized word power.

3. When listeners are trained on a master set of words, and when the subjective word parameters are determined from tests on that set, then articulation curves for various subsets can be predicted with good accuracy from these parameters, for the case where listeners receive no training on the subset. The error in such predictions is directly related to any inadvertent training (subset memorization) of the listeners, and can be kept small by taking reasonable precautions during subset tests.

4. To the extent that subset predictions are good estimates of the actual curves, one is able to obtain, by the methods described, subset curves that are directly comparable. For a large number of subsets, such comparable curves can be obtained by the word parameter method more easily, from the standpoint of time and stability of test conditions, than by conventional testing techniques.

5. The articulation curve can be shaped by choosing words on a basis of threshold and spread.

6. The word parameter approach provides a logical and consistent basis for examining the variation of intelligibility between individual words, between subsets, and between listeners.

7. Certain aspects of word-set intelligibility revealed by this study merit further investigation. One such aspect is the possibility of constructing a general method for synthesizing articulation curves having a specified shape; another is the possibility of quantitatively predicting the changes in articulation curves which result from subset training.

A P P E N D I C E S

APPENDIX A

PREPARATION OF TAPES

All recordings, both original and transcriptions, were made with Ampex 354-2 recorders at 7.5 inches per second. Of the two recorder channels, one was always used for speech, the other for noise; these two signals were recorded on, or played back from, two adjacent tracks on the quarter-inch magnetic tape. A "low-print" tape (Scotch 138) having 1.5 mil polyester backing was used to reduce "print-through" and insure dimensional stability. Shielded two-wire 600 ohm unbalanced lines were used in making all recordings and transcriptions except the original speech recordings; for these, the input was through a shielded unbalanced 200 ohm line. The 600-ohm impedance level was maintained by using precision resistors as external loads, and by bridging these lines only with high-impedance devices such as vacuum-tube voltmeters and crystal headphones. Recorders were aligned, adjusted, and calibrated to factory specifications before use, and were periodically maintained and calibrated during the research. Record and/or reproduce levels were set to standard values each time a recorder was used, such that the record-reproduce sequence resulted in a played-back replica of the original recorded signal, i.e, the recorder operated as a one-to-one transducer with any desired time lag between input and output. The output calibration was such that a standard 1000 cycle tone recorded on a tape played back at + 4 dbm; this corresponded to 0 VU on the output meter, a level not exceeded by more than 1.5 VU for any word peak.

Certain precautions were taken to minimize variations and distortions in the recorded signals and to assure a high degree of repeatability on playback, as follows:

1. All recording equipment was operated in air-conditioned rooms and a minimum of one hour warm-up time was required before use.

2. The erase, record, and playback heads were cleaned regularly with solvent, in order to prevent accumulation of foreign matter, such as tape oxide coating particles, in the air gaps. All parts of the tape path, including guides, rollers, and capstan, were kept scrupulously clean.

3. The recorder head assembly was demagnetized periodically.

4. Final measurements or transcriptions were not made from any tapes until they had been "worn in" by repeated playing.

5. All tapes were carefully bulk-erased before initial use. Virgin tape was used for all masters and copies.

6. All tape re-winding was done at low speed (7.5 ips) to prevent "print through."

7. All recorded tapes were protected from stray magnetic fields by storage in demagnetized steel containers.

8. All signals were monitored aurally and visually during recording, so as to catch any "drop-outs" due to non-uniform oxide coating or to the momentary lodging of an oxide particle in the head gap.

9. Extraneous "clicks" and other noise transients, recorded on the tape due to stopping and re-starting the recorder during a run, were avoided.

10. Tubes, capacitors, and resistors in the recorder amplifiers were replaced as they became noisy.

With the above precautions, a high degree of repeatability was obtained on playback, even with second-generation copies. It is estimated that recorded signal amplitudes were reproducible within ± 0.75 db, and that, within these limits, multiple copies of a master tape could be made which were essentially identical.

A 29-minute noise tape, used later to provide a reproducible noise for recording on one track of the master tapes, was recorded using the gaussian white noise output of a General Radio 1390-A noise generator. The output of the noise generator was passed through a low-pass filter (UTC LML 8000) having an eight kc cutoff frequency before being applied to the recorder input. A millisecond relay was used in conjunction with the timer and energy meter to sample and measure word-length sections of the noise waveform, with the sample time being determined by manually operating a switch. The timer measured the sample duration and operated the relay, while the energy meter measured the energy of the sample. In this way the power (energy divided by duration) of word length noise samples could be monitored at three points: the recorder input (filter output), the "record output" (an amplified version of recorder input after passing through the record level control and most of the record electronics), and the "playback output" (an amplified version of the signal played back from the tape). In addition, a Ballantine Model 320 "true rms" voltmeter was available for measuring the rms noise at these points. During a trial run to determine the noise generator setting for the desired noise power p_n, it was noted that p_n had both a short-time variation about some mean value as well as a long time drift over a 30-minute period. These drifts were stabilized by a four-hour warmup of

equipment. The noise generator output was then set to give approximately the same noise power as possessed, on the average, by the test words, and the recording was started. Measurements of p_n and rms noise voltage were made continuously at the recorder input and playback output, at average rates of three per minute for p_n and once every three minutes for rms voltage. By recording under continuously monitored conditions, subsequent power checks and possible re-recording of the tape were eliminated. From the recorded values, and from subsequent monitoring of the tape with headphones and VU meter, it was possible to select an 11-minute section of the tape which was free of drop-outs and whose long-time average p_n drifted by only one per cent. This section of the noise tape was used to transcribe noise to all master tapes.

The master tapes, on which word and noise power measurements were made and from which all listener tapes were copied, were made in three steps:

1. Recording of words on speech track.

2. Transcription of noise (from the noise tape) onto the noise track.

3. Recording of auditory cues for the listeners.

Initially, all 1000 of the Harvard PB words were recorded, although subsequently only 40 of these were selected for testing. Each of the twenty 50-word lists was recorded on a separate master tape. Four trained speakers (radio announcers) were auditioned before making a final selection of speaker. After instruction, familiarization with the words, and some practice, each speaker recorded the words of PB 1. After monitoring these recordings, the speaker having the best combination of

voice quality, pronunciation, and ability to sustain a constant vocal effort, was selected. This speaker then spent several hours familiarizing himself with all the PB words and, after being instructed to use constant vocal effort, in practicing word pronunciation. The speaker was instructed to use fairly crisp pronunciation without any marked rise or fall in pitch. He was also warned against clipped (unnaturally shortened) words and against allowing his effort at distinctly pronouncing each sound within a word to lengthen the word unnaturally.

The lists were recorded in a broadcast studio with the speaker reading the words from a printed list at 10-second intervals, the timing being determined by visual cue from the equipment operator. The microphone, a low-impedance dynamic type (Shure model 51), was positioned about eight inches in front of, and about two inches below, the speaker's mouth. Great care was taken to minimize the recording of extraneous noises on the tape, particularly during the 9.5-second interval between words (this later permitted the use of an extremely small voltage threshold setting in the timer). All doors to the sound-proofed studio were kept closed, and a large screen, made of sound-absorbent material, was used to partially enclose the recording equipment and operator and to separate them from the speaker, thus preventing tape transport rumble and motor noise from reaching the microphone. This highly anechoic environment reduced reverberation and microphonic effects to a minimum. Hum pickup was minimized by proper polarization of ac line cord connectors. All but one of the fluorescent light fixtures was turned off to minimize pick-up of a "crackling" noise traced to this source. Air noise was reduced by turning off the air conditioning equipment, and the

microphone was positioned so that its directional pattern discriminated against equipment noise.

Immediately after each word, the "head gate" of the recorder was opened by the operator and then reclosed just prior to the following word, thus temporarily destroying the tape-to-head contact and preventing any signal from being recorded in the inter-word interval. This prevented noises made by the speaker, such as from breathing or swallowing, from being recorded. Recordings were made in the evening or at night, when few people were in the building, thus minimizing noises from outside the studio. When, in spite of these precautions, aural monitoring of a tape revealed extraneous noises, the tape was discarded and a new one made. A vocal introduction followed by calibration tones was recorded at the start of each master tape. The speaker was required to adjust his vocal effort to give a standard VU meter deflection on several fixed test words before reading a word list, and was required to record as "practice words" the first five words of PB 1 near the start of each tape. The tapes were later monitored for correctness of word spacing, conformity with the list, voice quality, and pronunciation. Tapes lacking in any respect were discarded, and new ones made.

Next, a selected portion of the noise tape was transcribed to the noise track of each master tape. The noise crosstalk into the speech channel was observed and found to be below the ambient room noise recorded during words on the speech track. After monitoring the noise track of each master, those masters having no "drop-outs" near or during a word were selected for later transcription onto listener tapes. Finally, audible "cues" were recorded in the form of a short (0.5 second) burst

of two kc tone prior to each word and at a level of -7 dbm. The nominal cue-to-word spacing was three seconds; most cues were within 0.5 second of this position and no cue position was in error by more than one second. Cues were placed on the noise track, rather than on the speech track, for two reasons: cues on the speech track would have interfered with later power measurements, and cues on this track also would have become inaudible as the speech was attenuated during listening tests. The cues were recorded by manually switching the record head in such a way that the noise was effectively suppressed during each cue, thus providing a constant-amplitude, noise-free cue to alert the listeners for the following word. The final result of these procedures was a set of 20 master tapes, each having the recording sequence shown in Figure 1A.

Since it was planned to present to the listeners many different orderings of a given word set, it was not feasible to use the master tapes for listening tests. For this reason, and also to preserve the

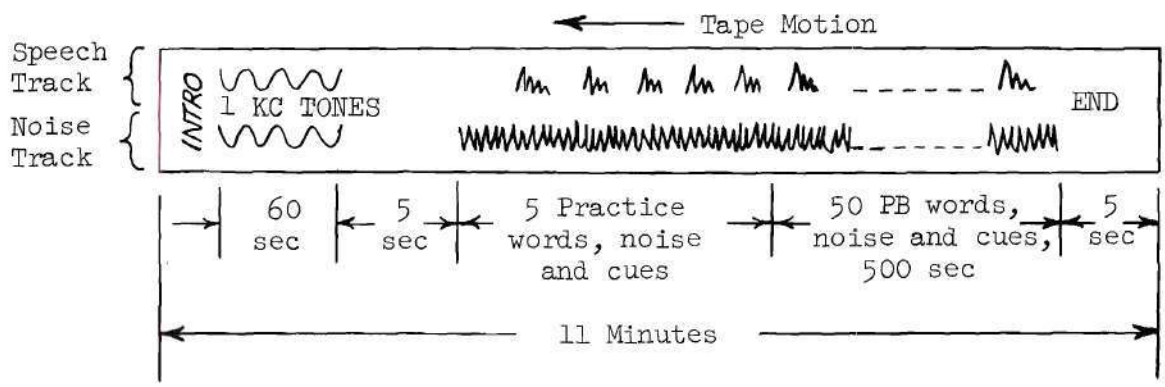


Figure 1A. Typical Master Tape.

masters, a large number of copies was transcribed from the masters. The uniformity of these copies was assessed by monitoring the noise track of 48 such copies with a VU meter. The results, on copies from four masters, revealed only ±0.5 db variation in average noise level from copy to copy (from the same master), and only ±0.1 db between the "copy average" (average noise level over all copies of a given master) for different masters. The extreme values of average noise level over all copies were only 1 db apart. The major cause of such variations was non-uniformity of the tape oxide coating, although errors in setting playback gain also played a part. Not only were there coating variations from tape to tape, but also long-time variations from end to end of a given tape. The largest observed result of the latter variation was 1.75 db end-to-end, with only 10 per cent of the tapes exhibiting end-to-end variations of more than 1 db. The results of these measurements were used as a partial basis for choosing the best tapes, with the result that maximum error in noise levels during listening tests was estimated to be ±0.75 db, with most errors being less than ±0.5 db.

All copies were monitored with phones and VU meter to screen out any tapes containing "drop-outs." On the basis of this test, together with the noise-uniformity tests just described, the 8 best copies of PB 3 and PB 20 were selected, out of the 38 available, for use in the listening tests. The final step in preparing the listener tapes was to cut the selected copies into strips, with one word on each strip, and to assemble eight 40-word tapes by splicing the strips together in random orders. Each such tape contained 39 words from PB 3 (all from the same copy), and 1 word from PB 20, i.e., each listener tape contained the master word

set listed in Table 1. Multiple transcriptions of the five "practice words" were made from one of the masters, and one of these spliced onto the start of each listener tape. Each strip forming part of a tape was identified as to word, copy number, and direction of play by visible colored markings, while each complete tape was identified by markings on its leader and trailer, as well as on the storage can. No calibration tones were placed on the listener tapes. Instead, several 10-second recordings of a calibration tone were made and formed into loops, using the same recorder, tape type, and level settings as were used in making the copies. Three of these "tone loops" having essentially identical tone levels were later used to calibrate the play-back machine for the listener tests, the playback machine having been modified so as to play either a tape loop or reel of tape, as desired. This approach, wherein standardized tone loops could be used for transfer calibration between record and playback machines, was considered more desirable than placing tones on each tape, particularly since the playback recorder needed calibrating only about once every hour.

The 20-word subset tapes, used after the master-set tests had been completed, were made by cutting the 40-word tapes into strips again, and reassembling these to form the A and B subset tapes, using the same procedures outlined above. Later, the A and B tapes were re-cut into strips, and reassembled to form G and H subset tapes. The subset tapes differed from master-set tapes in that only two practice words were used to precede the test words, and in the length of the strips (i.e., time interval between words). Due to the cut-and-splice process, which was also used in the continuous re-randomization program carried on during

the tests, a continuous reduction in strip-length occurred as a result of removing approximately 1.5 inches of tape during each splicing operation. Since this seemed to have no effect on listening (the original 10-second word spacing was greater than required by the listeners), all strips were shortened to 60 inches (8-second word spacing) before assembling the subset tapes.

The splices in the listener tapes caused an audible "click" and momentary interruption of the noise on playback, but, since these minor discontinuities occurred at regular intervals and well away from the cues or words, they caused no difficulty. The problem of storing tape strips prior to assembly was solved by a portable wood storage rack, assembled with non-magnetic screws and provided with labelled wooden pegs on which the strips were hung. This storage rack was kept well away from strong magnetic fields.

APPENDIX B

ARTICULATION TESTING PROCEDURES

The general nature of the articulation tests and test material has already been described, as have the various tapes used during the tests. The two variable test parameters of principal interest were the signal noise ratio and the composition of word sets; these were varied in a controlled manner to yield raw data in the form of stimulus-response pairs for each listener. This raw data was classified, tabulated, and reduced to yield desired quantities such as articulation scores or NMGF points.

As is well known such tests are affected by a large number of environmental, psychological, and procedural factors, only some of which are under control of the operator administering the tests. Some of these factors, along with measures taken to control them, are discussed below.

The listening team was composed of three male college students (JB, WN, and NS), all of whom had previous experience (ranging from 3 to 21 months) as members of articulation test teams. When the listeners were tested for hearing loss, using standard tone-audiometry techniques, the resulting audiograms showed normal hearing with no deficiencies. Intelligence of the listeners, based on grades in college courses, was judged to be above average, and neither personal interviews nor consultations with previous employers revealed any obvious mental or physical impediments to performance as members of a listening team.

The number of listeners is smaller than usual in articulation

testing. Aside from practical considerations such as availability of listeners, one factor of importance was the requirement of satisfactory statistical reliability of team scores, such reliability being directly related to the number of listeners. Since team scores were only one of several results to be obtained from the tests, and since number of repetitions and number of listeners are interchangeable as far as reliability of mean scores is concerned (25, p. 47), the three-listener team was considered adequate in view of the number of repetitions planned. That the resulting confidence in scores was satisfactory has already been illustrated in Chapter V. Another reason for not using a large number of listeners was the nature of the objective in making the tests. This objective was not to obtain results which would be "typical" for an "average" listener, but simply to establish the utility of the word threshold concept and to relate this to articulation scores. For this purpose, only one listener would have been sufficient, but the larger number of repetitions required, and the possibility of losing a listener before the tests were complete, argued against a single listener. In addition, the use of more than one listener permitted a comparison of listener responses to various words.

The primary objective of a "standardized" response, rather than a "typical" one, permitted a certain amount of flexibility in the type of response required of a listener. In particular, the commonly-used forced-choice response was not required, although, as discussed in Chapter IV, one of the listeners seemed to have made an almost perfect forced-choice response during the tests. The instructions to the listeners are reproduced in part below; as can be seen, they permit three types of response.

Listening Tests

1. No talking during runs.

2. Do not discuss tests with other listeners.

3. Do not change volume control without permission.

4. Report any equipment trouble.

5. Report any unusual sounds, tape defects, clues, or fatigue.

6. Use practice words to accustom yourself to the speech level.

Noise level will remain constant. Cues are about 3 seconds before words.

Responses to Tests

1. Initially, you will work with a set of 40 words, which should be memorized. You will not always know how many words to expect on a run.

2. At low SN ratios (low speech levels) you may not detect the presence of a word. Record a "dash" (-).

3. If you detect the presence of a word, but are unable to identify any specific sound, record a "plus" (+).

4. If you identify at least one sound, form the best possible estimate of the word and write it down before the next word is sent. If you are so uncertain that you have not decided by the time of the next cue, fill in a "plus" (+).

5. Do not go back and fill in spaces, make erasures, or mark through words, once the word is past.

Consistency

Consistency is the most important goal. Try to choose a level of listening effort which you will be able to maintain for two hours of tests. Overly intense concentration and strain will accelerate fatigue and cause

inconsistent responses. A certain amount of learning and improvement is expected, and is not inconsistent with constant listening effort.

Subjective Factors

Responses are greatly affected by variations in your mental and physical state. Report any unusual conditions such as colds, sickness, lack of sleep, and physical discomfort. Every effort will be made to develop a standard routine for tests and to avoid delays or variations of procedure.

Because of the above instructions concerning listener's responses, the articulation curves obtained are probably somewhat atypical, especially for low values of SN, but this is unimportant in view of the objectives of the tests. Listeners were not told any of the scores or test results, or what SN ratio was being used. The only identification of a run available to the listeners was a run number which they entered on the score sheet. The "volume control" referred to in the instructions was an individual level control (attenuator) placed at each listening station. The listeners were initially allowed to experiment with the adjustment of this attenuator until a comfortable listening level was obtained; the level was then fixed and no changes were allowed.

In order to maintain a fairly constant average sound level at the headphones, the noise level was held constant and the speech level was varied to obtain different signal/noise ratios. Since the noise was continuous, and since a word typically occupied only 0.5 second out of the 10-second word spacing, the speech level contributed very little to the average total (speech plus noise) sound level over the 10-second interval. The purpose in holding the average sound level approximately constant was

to avoid shifts in masked hearing threshold in the listeners. It has been shown (40) that essentially the same results are obtained by the above procedures as are obtained when the total (speech plus noise) level is held constant.

Some listener fatigue could not be avoided, of course, but studies have shown (3) that fatigue has little effect for as much as six hours per day of listening, when appropriate rest periods are provided. Furthermore, any residual fatigue effects were made consistent, from one day to the next, by testing at the same SN, in the same sequence, and at the same time within the test period, for each day of testing. In running a 40-word articulation curve, for example, the first point (lowest SN) was always obtained at the start of the test period, successive points at higher ratios were obtained at regular intervals during the period, and the final point (highest SN) was always obtained at the end of the period. Thus points for the same SN, but run on different days, should be affected in an identical manner by fatigue. The length of the test period was standardized at two hours per day, with a one minute break between eight minute runs and a fifteen minute break midway through the test period. Tests were run five or six days a week, over a period of about 12 weeks.

In addition to training, consistency of test procedures and conditions contributed to standardizing the listeners' performance. Tests were run in the afternoons, with listeners seated side-by-side at individual listening stations separated by movable baffles. Several techniques were used to minimize extraneous noises. The testing room itself was sound-deadened by acoustical treatment of the walls and ceiling, and

oversize foam rubber cushions were used on the headphones. A warning light over the door to the room prevented interruptions during tests. During most runs, the loudest noise audible in the testing room was that made by the recorder tape transport.

Each run was started by having the listeners fill in the heading of a standard score sheet, shown in Figure 1B. Minor modifications were made in the score sheet after training was completed, consisting principally of removing the numbers from the spaces provided for responses. Each run proceeded according to a standard sequence, as illustrated by the typical test log entry below:

Run	Date	Time	Monitor	Tape	Atten.
292	2-23-62	2:08		11118	26
293	2-23-62	2:16		11101	24
294	2-23-62	2:25		11114	22
295	2-23-62	2:32		11102	20
296	2-23-62	2:42		11103	18
297	2-23-62	3:01		11105	16
298	2-23-62	3:10		11117	14
299	2-23-62	3:18		11106	12
300	2-23-62	3:25		11104	10
301	2-23-62	3:34		11109	8

Test log entries were made by the test monitor. The first four items in each test log entry were also recorded by the listeners on their score sheets, along with the listener's initials and his headphone level. Score sheets were taken up as soon as completed. Grading of score sheets

GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING

Articulation Test Sheet

Run Number	Monitor	Score
Date	Listener	Scorer
Time	Level	Tape Number

1.		21.		
2.		22.		
3.		23.		
4.		24.		
5.		25.		
6.		26.		
7.		27.		
8.		28.		
9.		29.		
10.		30.		
11.		31.		
12.		32.		
13.		33.		
14.		34.		
15.		35.		
16.		36.		
17.		37.		
18.		38.		
19.		39.		
20.		40.		

Figure 1B. Listener Score Sheet.

was usually done by the monitor during the test period and was later re-checked after all tests had been completed. In addition to the above, the monitor also instructed the team, operated the equipment, announced rest periods, and collected and filed the score sheets. The monitor was responsible for calibrating the equipment, cleaning the recorder tape path and heads, and rewinding the tapes.

The tape recorder playback levels were set to standard values at the start of each test period and again midway through the test period, using the calibration tone loops described in Appendix A. After checking the speech and noise channels separately, the total (mixed) signal level was checked at the output of the mixer. Finally, the level at the eight-ohm line to the headphones was set at a standard level by means of the audio amplifier gain control.

The monitor was able to aurally and visually monitor the signals during a run by means of headphones, vacuum-tube voltmeters, and VU meters. After an initial familiarization period, each of the two monitors available was able to carry out tests at a rate of approximately 260 word transmissions per hour, including breaks, tape changing, and calibration.

It has been long known (1) and is widely recognized (24, 26) that the amount of experience and practice possessed by a listener has a significant effect on scores. This is known as "learning effect" or "practice effect" and is attributed to the increasing familiarity of the listener with the test items as well as his increasing ability, in the noise-masked case, to "hear through the noise." Another factor which results in higher-than-expected scores is memorization, i.e., the memori-

zing of the order-sequence in which test items are sent. This effect, which is observed only where the same sequence of items is transmitted many times (as occurs when using taped word lists), leads to situations where a sequence of several words is known immediately upon recognizing a key word-pair. The effects of learning are usually minimized by following a listener training program, while memorization is minimized by repeating a given word sequence as few times as possible. A third, and usually less important, effect is the listener's association of a word with some acoustic "clue" such as a sharp break in noise level near the word or perhaps an unusually long or short burst of tone used for cueing. In such cases the listener responds with the correct word because he recognizes the clue and not because he recognizes the word.

In spite of precautions, some memorization apparently took place during the subset tests, as discussed in Chapter IV. Auditory clues were minimized by using great care in producing the listener tapes (see Appendix A), so as to introduce as few anomalies into the tapes as possible. As far as could be determined by monitoring the tapes and questioning the listeners, all tapes were essentially identical, except for word order, and free of auditory clues. Only one or two such clues were reported by the listeners during the tests.

It has been observed (3), even when memorization effects and acoustical clues are absent, that practice results in considerable improvement in scores. This "learning" results from increasing familiarity with test procedures, speaker's voice, vocabulary, and with the way various words sound when masked by noise (the same word may sound different at different signal/noise ratios). These effects are commonly

exhibited by means of a "learning curve" such as that in Figure 2B. This curve, which was obtained for a single listener (NS) and fixed value of SN (-12.75 db), shows the improvement in score, for a 40-word test, with the number of repetitions of the test. This curve illustrates the typical high learning rate which occurs at the start of the tests and the gradual stabilization of scores as the curve reaches a plateau, i.e., as training is completed. Actual test results reported and used in this thesis were taken only after a training period of 21 repetitions, as shown on the curve. Since the learning process never ceases entirely, the experimenter's objective is to reach a point where residual learning is statistically insignificant compared to other factors affecting the scores. To this end, the training program described below was followed.

After initial familiarization and practice, the listeners ran several 1000-word curves, using all 20 PB lists and eight values of SN. The results of this test, which involved NS, JB, WN, and three additional listeners are reported elsewhere (9), and indicate generally typical performance. Following this, JB, WN, and NS were selected as members of the final listening team, and their training on the 40-word master list was begun. The listeners were given a copy of this master list and asked to memorize the words and their spelling. Several hours of verbal and written practice in repeating the word list followed, including practice in identifying list words from a general collection of PB words. The team then listened to four repetitions of a tape containing the master list but without noise masking, writing a response to each transmitted word before consulting a printed version of the word sequence on the tape.

Next, a series of training repetitions was begun, each repetition

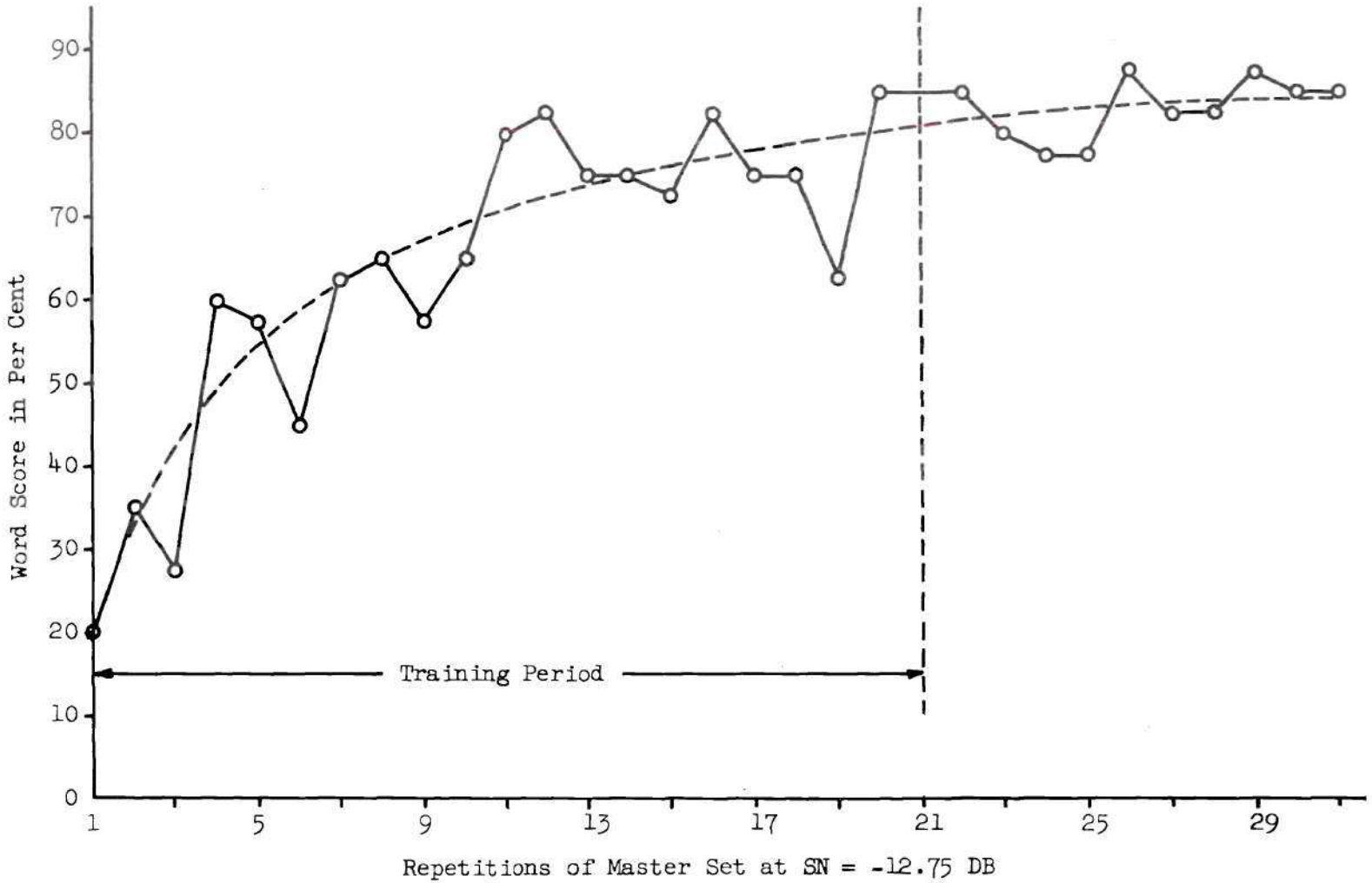


Figure 2B. Learning Curve for Listener NS.

consisting of a complete 10-point articulation curve for the master list. The words and procedures for these training runs were the same as for the final data runs, except that the SN increment was 3 db instead of the 2 db used in data runs. Ten randomized versions of the list were played at different signal/noise ratios for each repetition, with SN being varied (in equal steps) so as to cover a word score range of approximately 5 to 95 per cent. There was some initial experimentation to determine the SN range to be covered, and this range had to be changed from time to time as learning progressed. Initially, the range was -18.75 db to -0.75 db, but by the end of the training it had become -30.75 db to -6.75 db. With two exceptions, a repetition was made by increasing, in steps, the value of SN; the results of the two decreasing SN repetitions were not significantly different. Copies of the word list were available to the listeners during these runs, but were not used after about ten runs.

One way of exhibiting learning progress is to plot curves similar to Figure 2B for each listener and each value of SN. Such a set of curves (approximately 30 curves in the present case) would reveal learning at different rates for the various listeners, as well as the fact that learning rate is also a function of SN. That is, the superiority of one listener's learning rate at a given SN may be reversed at a different SN. Perhaps the most inclusive way of plotting a learning curve is to plot score per repetition, i.e., the team scores averaged over all values of SN used in the repetition, as a function of number of repetitions. This is difficult in the present case because of the variation in SN range which took place during training. As a compromise, scores were averaged over the listeners and over those values of SN common to all

training runs, with the results shown in Figure 3B. This team learning curve shows that learning rate had slowed considerably by the end of the training period. The number of practice hours (approximately 27) and repetitions (21) required for training was consistent with previously-reported results (40, p. 40) (25, p. 34).

As mentioned above, the learning rates vary in a complicated way from listener to listener, with SN, and with repetitions. A method utilizing analysis of variance techniques has been developed (25) which provides a more sophisticated estimate of the point at which learning becomes statistically insignificant. This method is based on the assumption that scores vary with listener, SN, repetition, and experimental error, and, in addition, with first-order interactions between these factors. This method was applied three times during training, the calculations being made on a digital computer (Burroughs 220). The analyses for repetitions 1 through 5 and 6 through 9 showed significant learning, as was expected. The third analysis, for repetitions 16 through 21, showed significantly lower scores on repetition 19 which were subsequently traced to low scores on two of the ten runs comprising that repetition. These runs being sequential, it was suspected that some temporary condition (probably excessive ambient room noise) caused the low scores. The fact that the scores were low rather than high argues against learning as the cause; consequently, this entire repetition was dropped from the analysis, with the result that repetition effect and all interactions involving it showed no significance. Thus, based on the arbitrary, though reasonable, criterion of five successive non-significant repetitions, the team was assumed to be fully trained after repetition

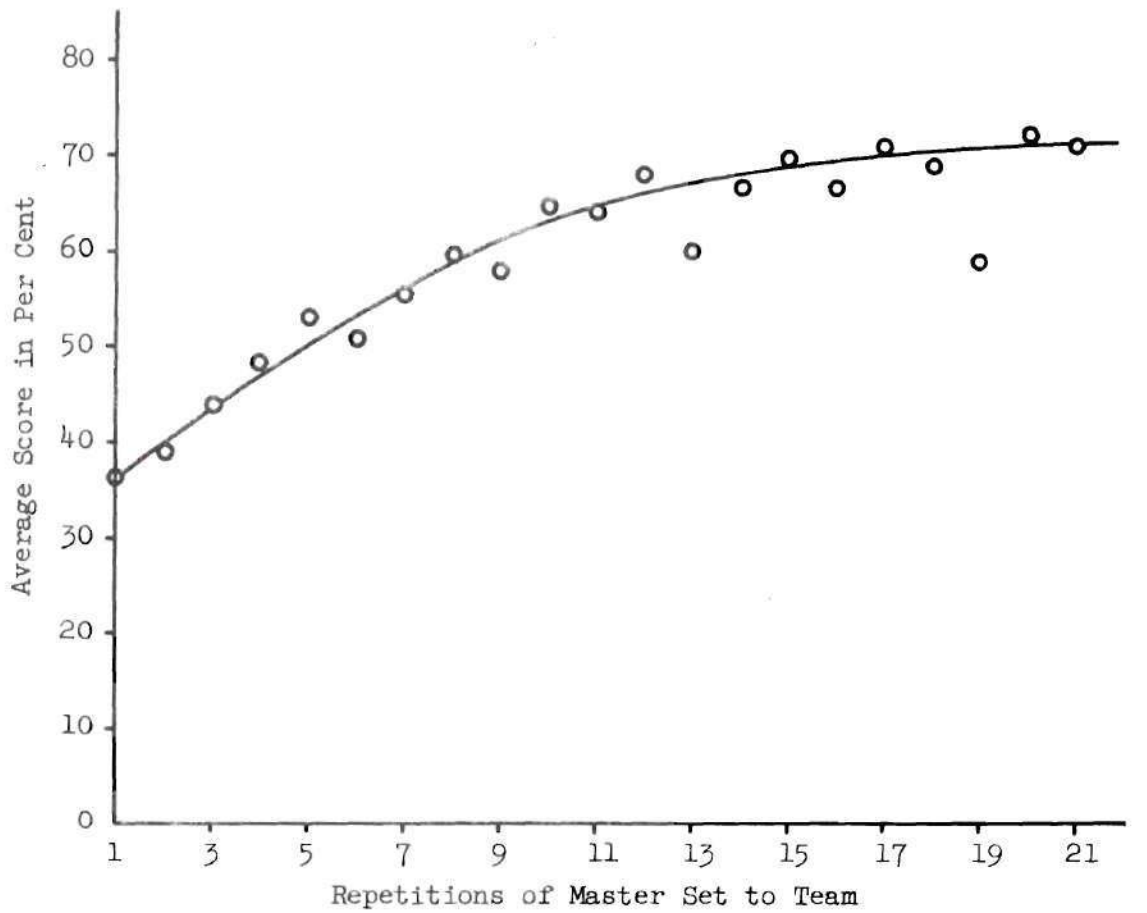


Figure 3B. Learning Curve for Team.

21. This method is superior to use of the learning curve in Figure 3B in that all eight SN ratios were used instead of the five used to plot Figure 3B.

A program of re-randomizing word sequences, which was carried on continuously during the tests, was accomplished by cutting the tapes into strips and reassembling these in a new order. Questioning of the listeners revealed that they had begun to remember a few two-word sequences by the end of the ninth repetition. Also, some listeners had begun to remember that certain words occurred at some rather vaguely-defined point in a sequence. For example, they might remember that the word "fig" occurred near the bottom of the first column of a score sheet for one of the tapes, but could not pinpoint it as being, say, the nineteenth word. In order for such memorization to be effective, the listeners must have some way of identifying a tape; this seemed to occur mainly through using the first word on the tape as a "key," an action easily frustrated by moving the first word to some new location in the tape. Because of this, and since a complete re-randomization of a tape was a lengthy and tedious process, most of the re-randomizing consisted of dividing a tape into two to four sections and shifting the relative positions of these sections, taking care to break up any two-word sequences known to have been memorized.

As a final precaution, the order in which tapes were presented during the data runs was varied. To avoid consistently playing a given tape at high signal/noise ratios where memorization is easiest, tests were arranged so that, in each sequence of eight repetitions, each tape was heard once and only once at each value of SN. This was accomplished for

the eight test values of SN by arranging the eight tapes in a Latin square (37, p. 77), as shown in Figure 4B.

In Figure 4B, the test tapes have been arbitrarily numbered 1 through 8. The eight columns (or eight rows, for that matter) furnish a set of eight tape sequences, i.e., eight repetitions, in which each tape is heard at each value of SN. Averaging scores over at least eight repetitions would tend to average out any differences in the tapes and to make the scores independent of any such differences.

8	6	1	3	4	2	5	7
7	3	8	4	1	5	6	2
6	8	2	5	7	4	3	1
5	4	3	7	2	1	8	6
4	1	5	6	3	7	2	8
3	2	7	1	6	8	4	5
2	7	6	8	5	3	1	4
1	5	4	2	8	6	7	3

Figure 4B. Latin Square of Tape Numbers.

APPENDIX C

DEFINITIONS OF SYMBOLS

1. $e^i(t)$ = Instantaneous word voltage waveform during interval T^i, for i^{th} word.
2. $e_n^i(t)$ = instantaneous noise voltage during the T^i-second interval associated with the i^{th} word.
3. $F_\beta(x)$ = $\text{Prob}\left[\beta \leq x\right]$ = distribution function for β.
4. $F_\xi(x)$ = $\text{Prob}\left[\xi \leq x\right]$ = distribution function for ξ.
5. $G^i(SN)$ = linear approximation to NMGF, expressed as a function of SN, for i^{th} word.
6. N = number of words in a set.
7. NMGF = noise-masked gain function.
8. $P^i = 10 \log \frac{p^i}{\overline{p}}$ = power of i^{th} word in db relative to $\overline{p}$.
9. $P_n^i = 10 \log \frac{p_n^i}{\overline{p}_n}$ = power of i^{th} noise segment in db relative to $\overline{p}_n$.
10. $\overline{p}_n = \frac{1}{N}\sum_{i=1}^{N} p_n^i$ = average noise power over set.
11. $\overline{p} = \frac{1}{N}\sum_{i=1}^{N} p^i$ = average word power over set.
12. $p^i = \frac{1}{T^i}\int_{T^i} \frac{1}{R}\left[e^i(t)\right]^2 dt = \frac{w^i}{T^i}$ = power of i^{th} word.

13. $p_n^i = \frac{1}{T^i}\int_{T^i} \frac{1}{R}\left[e_n^i(t)\right]^2 dt = \frac{w_n^i}{T^i}$ = power of T^i-second segment of noise coincident with i^{th} word.

14. R = Resistive impedance across which a speech or noise waveform is produced at some standard measurement point.

15. r = linear correlation coefficient for points fitted with least-squares line.

16. $SN = 10 \log \frac{\bar{p}}{\bar{p}_n}$ = group signal/noise ratio, in db, for entire word-set.

17. $SN^i = 10 \log sn^i$ = signal/noise ratio for i^{th} word in db.

18. $sn^i = \frac{p^i}{p_n^i}$ = signal/noise ratio for i^{th} word.

19. T^i = time duration of i^{th} word.

20. w^i = energy of i^{th} word.

21. w_n^i = energy of noise which masks i^{th} word.

22. X = "gain" from input to output of attenuator. Note that $|X|$ is the number of db of attenuation and that $X = SN^i - \left[SN^i\right]_{measured} \leq 0$.

23. $|X_o|$ = attenuation for zero word score on a linear approximation to an NMGF, i.e., the X-intercept of a $G^i(X)$ function.

24. $|X_{50}|$ = attenuation for 0.5 word score (50 per cent score).

25. α^i = threshold value of SN^i, for i^{th} word.

26. $\beta^i = \alpha^i - P^i + P_n^i$ = threshold value of SN for i^{th} word.

27. Δ^i = spread in db of i^{th} NMGF or of its linear approximation, from 0.0 to 1.0 fractional score.

28. ξ^i = $\alpha^i - SN^i$ = value of intelligibility variable ξ for i^{th} word.

BIBLIOGRAPHY

1. Fletcher, Harvey, and Steinberg, J. C., "Articulation Testing Methods," Bell System Technical Journal, Vol. VIII, Oct. 1929.

2. Fletcher, Harvey, and Galt, Rogers H., "The Perception of Speech and Its Relation to Telephony," Journal of the Acoustical Society of America, 22, No. 2, March 1950, p. 89.

3. Egan, J. P., "Articulation Testing Methods-II," OSRD Report 3802, Harvard Psycho-Acoustics Laboratory (1944).

4. Hawley, Mones E., "Articulation Testing and Standards for Articulation Tests," paper given at 48th meeting of Acoust. Soc. Am., Austin, Tex., 1954.

5. Hirsh, I. J., and others, "Development of Materials for Speech Audiometry," Journal of Speech and Hearing Disorders, 17 (1952), pp. 321-337.

6. Hawthorne, G. B., Jones, W. B., and others, "Performance of Communication Systems in the Presence of Interference," Final Report, Vol. I, Contract AF 30(602)-1789, Engineering Experiment Station, Georgia Tech, 1 June 1959.

7. Licklider, J. C. R., and Pollack, I., "Effects of Differentiation, Integration, and Infinite Peak Clipping upon the Intelligibility of Speech," Jour. Acoust. Soc. Am. 20 (1948), p. 42.

8. Ferrell, P. J., "Constant Amplitude Speech," RADC Technical Note 57-317, Rome Air Development Center (1957).

9. Robertson, Douglas W., "A Comparison of the Procedures and Results of Intelligibility Tests for a Number of Interference Conditions," Technical Memorandum No. X003-10, Electromagnetic Compatibility Analysis Center, U. S. Naval Engineering Experiment Station, Annapolis, Md., April 16, 1962.

10. Kryter, Karl D., "Speech Communication in Noise," AFCRC-TR-54-52 (Technical Report), Air Force Cambridge Research Center, Washington, D. C., May 1955.

11. Kryter, Karl D., "Effects of Ear Protective Devices on the Intelligibility of Speech in Noise," Jour. Acoust. Soc. Am. 18, No. 2, October 1946, p. 413.

12. Licklider, J. C. R., "The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise," Jour. Acoust. Soc. Am. 20, No. 2, March 1948, p. 150.

13. Pollack, Irwin, "Effects of High Pass and Low Pass Filtering on the Intelligibility of Speech in Noise," Jour. Acoust. Soc. Am. 20, No. 3, May 1948, p. 259.

14. Stevens, S. S., Miller, Joseph, and Truscott, Ida, "The Masking of Speech by Sine Waves, Square Waves, and Regular and Modulated Pulses," Jour. Acoust. Soc. Am. 18, No. 2, Oct. 1946, p. 418.

15. Hawkins, J. E., Jr. and Stevens, S. S., "The Masking of Pure Tones and of Speech by White Noise," Jour. Acoust. Soc. Am. 22, No. 1, Jan. 1950, p. 6.

16. Miller, G. A., Language and Communication, McGraw-Hill Book Co., New York, 1951.

17. Siegenthaler, B. M., and Gunn, G. H., "Factors Associated with Help Obtained from Individual Hearing Aids," Journal of Speech and Hearing Disorders 17 (1952), pp. 338-347.

18. Hudgins, C. V., and others, "The Development of Recorded Auditory Tests for Measuring Hearing Loss for Speech," The Laryngoscope 57 (1947), pp. 57-89.

19. Lewy, Alfred, and others, "Functional Examination of Hearing," Archives of Otolaryngology 50 (1949), p. 336.

20. Harris, J. D., "Some Suggestions for Speech Reception Testing," Archives of Otolaryngology 50 (1949), pp. 388-405.

21. Siegenthaler, Bruce B., and Hardick, Edward, "Intelligibility Scores Using Various Phonetically Balanced Word Lists," Pennsylvania Speech Annual, 1959.

22. Curry, Fay, and Hutton, "Experimental Study of the Relative Intelligibility of Alphabet Letters," Jour. Acoust. Soc. Am. 32, No. 9, Sept. 1960, p. 1151.

23. Clarke, Frank R., "Constant-Ratio Rule for Confusion Matrices in Speech Communication," Jour. Acoust. Soc. Am. 29, June 1957, p. 715.

24. Beranek, Leo L., Acoustics, McGraw-Hill Book Co., New York, 1954.

25. Stuckey, Charles W., "Development of a Method to Evaluate Learning in Articulation Testing," Technical Note No. 1, Georgia Tech Engineering Experiment Station (Project No. A-483), Sept. 1960.

26. Fletcher, Harvey, Speech and Hearing in Communication, D. Van Nostrand Co., New York, 1953.

27. Kryter, Karl D., "On Predicting the Intelligibility of Speech from Acoustical Measures," Journal of Speech and Hearing Disorders 21 (1956), p. 208.

28. French, N. R., and Steinberg, J. C., "Factors Governing the Intelligibility of Speech Sounds," Jour. Acoust. Soc. Am. 19, (1947) pp. 90-119.

29. Beranek, L. L., "Noise Control in Office and Factory Spaces," Transactions of Chemical Engineering Conference, Industrial Hygiene Foundation of America, No. 15, 1950.

30. Strasberg, M., "Criteria for Setting Airborne Noise Level Limits in Shipboard Spaces," Report No. 371-N-12, Dept. of the Navy, Bureau of Ships (1952).

31. Tkachenko, A. D., "Tonal Method for Determining the Intelligibility of Speech Transmitted by Communications Channels," Soviet Physics-Acoustics, 1, No. 2 (January-June, 1955).

32. Pickett, J. M., and Kryter, Karl D., "Prediction of Speech Intelligibility in Noise," paper contributed at 47th meeting of Acoust. Soc. Am., New York, June 1954.

33. Stevens, S. S. and Davis, Hallowell, Hearing, John Wiley and Sons, New York, 1938.

34. Curry, E. T., "An Experimental Study of the Relative Identification Thresholds of Nine American Vowels," Speech Monographs, XVII, March 1950.

35. Dunn, H. K., and White, S. D., "Statistical Measurements on Conversational Speech," Jour. Acoust. Soc. Am. 11, Jan. 1940, p. 278.

36. Sherwin, Kodman, and others, "Detection of Signals in Noise: a Comparison between the Human Detector and an Electronic Detector," Jour. Acoust. Soc. Am., 28, No. 4, July 1956, p. 617.

37. Freund, Livermore, and Miller, Manual of Experimental Statistics, Prentice-Hall, Englewood Cliffs, N. J., 1960.

38. Mosteller, Rourke, and Thomas, Probability with Statistical Applications, Addison-Wesley Publishing Co., Reading, Mass., 1961.

39. Feller, W., An Introduction to Probability Theory and Its Applications, John Wiley and Sons, New York, 1950.

40. Robertson, D. W., and Stuckey, C. W., "Engineering Investigation and Study of Automatic Voice Intelligibility Computation Methods," Final Report, Project A-483, Ga. Tech Engineering Experiment Station, Atlanta, Ga., Feb. 1, 1962.

VITA

George Boltz Hawthorne Jr. was born in Bainbridge, Georgia, on August 15, 1928. He is the son of George B. and Martha Rudisill Hawthorne. He was married in 1953 to Miss Ethel M. Lynch of Augusta, Georgia, and has two children.

He attended public schools in Bainbridge, Georgia and Sylvester, Georgia, and was graduated from the Sylvester High School as valedictorian of his class in 1945. He attended the Georgia Institute of Technology, where he graduated with highest honor in 1951, receiving the degree of Bachelor of Electrical Engineering. In 1956, he received the degree of Master of Science in Electrical Engineering from the same school.

His experience includes seven years as Assistant Research Engineer, Research Engineer, and Research Associate in the Engineering Experiment Station, and five years as Lecturer and Assistant Professor in the School of Electrical Engineering, all at the Georgia Institute of Technology.

He is a member of Phi Eta Sigma, Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi. He is also a member of the Institute of Radio Engineers and of its Professional Group on Information Theory, the American Institute of Electrical Engineers, and the Institute of Mathematical Statistics. He is a Registered Professional Engineer in the State of Georgia.